DOCUMENT RESUME

ED 171 792 TM 009 393

TITLE	Proceedings: Annual Conference of the Military Testing Association (20th, Oklahoma City, Oklahoma, October 3-November 3, 1978). Volume 1 [and] Volume II.
INSTITUTION	Coast Guard Inst., Oklahoma City, Okla.
PUB DATE	78
NOTE	1,473p.; Not available in paper copy due to marginal legibility of original document
EDRS PRICE	MF12 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Educational Programs; Females; Job Analysis; *Military Personnel; Occupational Surveys; *Occupational Tests; *Performance Tests; *Personnel Evaluation; *Personnel Management; Personnel Selection; Rating Scales; Recruitment; Simulation; Test Construction; Testing; Test Validity
IDENTIFIERS	*Military Testing Association

ABSTRACT

Papers presented at the Military Testing Association Conference focused on technical information and experience in the area of personnel management. Papers were contributed by representatives of American and foreign business, educational, and military communities. Topics for the 17 sections comprising the conference were: (1) Occupational surveys; (2) occupational task analysis; (3) women in the armed services; (4) general problems; (5) personnel appraisal; (6) the use of rating scales; (7) personnel selection; (8) methods of determining personnel availability; (9) validation and prediction of job performance; (10) statistical and measurement methodologies; (11) testing techniques and technology; (12) test items; (13) training programs and problems; (14) instructional evaluation and test development; (15) performance feedback; (16) performance measurement; and (17) simulators and simulation. The report of the steering committee, the by-laws of the Military Testing Association, and an index of authors and list of conferees are appended. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED171792

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

20TH ANNUAL CONFERENCE

OF THE

MILITARY TESTING ASSOCIATION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. J. Furst
U.S. Coast Guard

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

PROCEEDINGS

COORDINATED BY
UNITED STATES
COAST GUARD INSTITUTE

VOLUME I

OKLAHOMA CITY, OKLAHOMA
30 OCTOBER – 3 NOVEMBER 1978

TM009 398

25

PROCEEDINGS

20th Annual Conference
of the
Military Testing Association

Coordinated By:

U. S. COAST GUARD INSTITUTE
P. O. Substation 18
Oklahoma City, Oklahoma 73169

HILTON INN WEST
Oklahoma City, Oklahoma
30 October - 3 November 1978

Volume I

FOREWORD

The papers presented at the Twentieth Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of armed service.

TABLE OF CONTENTS

VOLUME I

Foreword. i

Keynote Address
"Quality of Life . . . . RADM W. H. Stewart. xii

Official Program. 1

SECTION 1 . . . . .OCCUPATIONAL SURVEYS 15

"Civilian Ground Safety Officer Job and Training Requirements Survey" . . . . Douglas K. Cowan. 16

"Determining the Training Requirements of United States Coast Guard Warrant and Commissioned Officer Billets" . . . . . J. W. Cunningham. 28
D. W. Drewes

"Evaluating the Army Occupational Survey Program Methodology: Answer Booklets, Questionnaire Length, and Population Coverage" . . . . Dr. Eugene M. Burns 51

"The Use of Job Satisfaction Data in the Occupational Survey Program". . . . . CAPT John Olivo 65
CAPT Elena J. Weber

"General Overview and Initial Findings of the Project on Job Satisfaction and Retention of U.S. Army Enlisted Personnel". . . . . Dr. Lawrence A. Goldman . . 75
Darrell A. Worstine
Cedella J. Bonette

SECTION 2 . . . . OCCUPATIONAL - TASK ANALYSIS. 111

"Execution of Large Occupational Analysis of the Royal Navy's Operations Branch". . . C. D. Beel. 112

"A Strategy for Task Analysis and Criterion Definition Based on Multidimensional Scaling". . . . . LCOL G. M. Rampton. 132

"Obstacles to and Incentives for Standardization of Task Analysis Procedures" . . . . Dr. Robert W. Stephenson. . 188
Hendrick W. Ruck

"Task Analysis: Destination or Journey" . . . Dr. Melvin D. Montemerlo. . 199
Dr. Francis M. Aversano

"Four Fundamental Criteria for Describing
the Tasks of an Occupational Specialty" . . Dr. Walter E. Driskell. . . 204
CAPT Frank C. Gentner

"Two Applications of Occupational Survey
Data in Making Training Decisions". . . . . CAPT David S. Vaughan . . . 213

"The Stability Over Time of Air Force
Enlisted Career Ladders as Observed in
Occupational Survey Reports". . . . . Dr. Walter E. Driskill. . . 228
CAPT Frederick B. Bower, Jr.

"The Collection and Prediction of
Training Emphasis Ratings for Curriculum
Development". . . . . Hendrick W. Ruck. 242
Nancy A. Thompson
David C. Thomson

"Data Base To Determination of Training
Content: A Manageable Solution". . . . . . D. Douglass Davis 258

"Using the Computer to Build the Task
Inventory". . . . . Thomas M. Ansbro. 263

"Systematic Instructional Validation
Through Testing". . . . . Dr. Marjorie A. Kuenz . . . 275
Fred C. Roberts

"Scheduling Formal School Training to
Maximize Cost Effectiveness". . . . . Doug Goodgame 286

"Methods for Determining Safety Training
Priorities for Job Tasks" . . . . Ms. Nancy Thompson. 296
Hendrick W. Ruck

"Methods for Collecting and Analyzing
Task Analysis Data" . . . . Dr. A. John Eschenbrenner . 314
Philip B. DeVries
Hendrick W. Ruck

"Methodology for Selection and Training
of Artillery Forward Observers Job
Analysis" . . . . Dr. John B. Mocharnuk . . . 324
Ruth Ann Marco

"Observer Self-Location Ability and its
Relationship to Cognitive Orientation
Skills" . . . . Dr. John R. Milligan. . . . 333
Dr. Raymond O. Waldkoetter

"Job Analysis in the US Army Medical Training Environment" . . . . J. S. Tartell 354

"CODAP: A New Modular Approach to Occupational Analysis . . . . Michael C. Thew 362
Johnny J. Weissmuller

"Occupational Analysis for Field Grade Army Officers". . . . . Ms. Sally J. Van Nostrand . 373
M. Reid Wallis

"A Technique for Selecting Electronic Specialties for Consolidation". . . . . . . Hendrick W. Ruck. 385

SECTION 3 . . . . WOMEN IN THE ARMED SERVICES 395

"Differential Field Assignment Patterns for Male and Female Soldiers" . . . . Dr. L. W. Oliver. 396
Ms. Nehama Babin

"The Premature Attrition Rate of Navy Female Enlistees. . . . . Gerry L. Wilcove. 420
Patricia J. Thomas
Constance Blankenship

"Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Description of Leader Behavior" . . . . MAJ Jerome Adams, Ph.D. . . 434
Dr. Jack M. Hicks

"Female Utilization in Non-Traditional Areas". . . . . Joseph A. Bergmann. 444
Raymond E. Christal

SECTION 4 . . . . GENERAL 462

"Strain by Prolonged Duty Hours and Problems as to Mobility of Soldiers--As Seen by Federal Armed Forces Association". . . . . COL Han-Erich Seuberlich. . 463

"Computer Assisted Reference Locator (CARL) System: An Overview". . . . . William A. Sands. 470

SECTION 5 . . . . .PERSONNEL APPRAISAL. 487

"Quality of ROTC Accessions to the Army Officer Corps". . . . . Dr. Arthur C. F. Gilbert. . 488
Dr. John I. Weldon
Dr. Richard S. Wellins

"Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB)". . . . John J. Mathews 494
Dr. Lonnie D. Valentine, Jr.
MAJ Wayne S. Sellman

SECTION 6 . . . . .USING RATING SCALES. 507

"The Content Issue in Performance Appraisal Ratings". . . . . CAPT Randy H. Massey. . . . 508
C. J. Mullins
James A. Earles

"Differential Responses on Alternately Anchored Job Rating Scales" . . . . LT COL Jimmy L. Mitchell. . 525

"Sample Size and Stability of Task Analysis Inventory Response Scales" . . . . Dr. John J. Pass. 537
David W. Robertson

"Benchmark Scales for Collecting Task Training Factor Data" . . . . David C. Thomson. 556
Kenneth Goody

SECTION 7 . . . . .PERSONNEL SELECTION. 565

"Weighted Selection System for AFROTC Applicants--Perspective After Second Year of Use". . . . . LT COL David K. Jackson . . 566
M. Meriwether Gordon

"The Defense Language Aptitude Battery (DLAB)" . . . . Robert G. Henderson 574

"Monte Carlo Computer Programs for Simulating Selection Decisions from Personnel Tests". . . . . John W. Thain 586

SECTION 8 . . . . .METHODS OF DETERMINING PERSONNEL AVAILABILITY. 601

"PAM: A Methodology for Predicting Air Force Personnel Availability" . . . . Harry A. Baran. 602
Duncan L. Dieterly
Andrew J. Czuchry
John C. Goclowski
Fredric F. Phillips
Stuart E. Peskoe
Anthony J. Lofaso

SYMPOSIUM:

Methodology for Mobilization Population Inventory . . . . .Chairman: Dr. Jack M. Hicks 632

"Some Implications of Commercial Test Normings for Mobilization Surveys". . . . . R. F. Boldt 633

"Measuring the Military Base Population of the 1980's" . . . . Dr. M. A. Fischl. 640

"Development of a Mobilization Population Inventory Using Existing ASVAB Data Banks". . . . . . George V. Rux 645
Dr. William W. Graham, Jr.

"Air Force Experience with PROJECT TALENT" . . . . Dr. Lonnie D. Valentine, Jr. 671

END OF VOLUME I

VOLUME II

SECTION 9 . . . . VALIDATION - PREDICTION 676

"The Impact of Valid Selection Procedures on Workforce Productivity" . . . Frank L. Schmidt 677
John E. Hunter
Robert C. McKenzie
Tressie W. Muldrow

"Job Performance of USAF Bypassed Specialists". . . . . CAPT William H. Cummings. . 724
CAPT David S. Vaughan

"Analysis of Heavy Equipment Operator Jobs" . . . . Dr. Sidney A. Fine. 734
Howard C. Olson
David C. Myers
Margarette C. Jennings

"Predictive Utility of the Officer Evaluation Battery (OEB)" . . . . Dr. Arthur C. F. Gilbert. . 753

"Assessment Center Variables as Predictors of On-Job Performance Characteristics". . . . . Charles H. Cory 761

"Using an Assessment Center to Predict Leadership Course Performance of Army Officers and NCOs" . . . . Dr. Frederick N. Dyer . . . 779
Richard E. Hilligoss

"Validity of Associate Ratings of Performance Potential by Army Aviators" . . . . Robert F. Eastman 823
Marie Leger

"Performance Test Objectivity: Comparison of Interrater Reliabilities of Three Observation Formats". . . . . Wi1liam A. Nugent 831
G. J. Laabs

"Prediction of Field Artillery Officer Performance". . . . . Dr. Arthur C. F. Gilbert. . 839
Dr. Raymond O. Waldkoetter
Anthony E. Castelnovo

SYMPOSIUM:

Innovative Test Validation Strategies . . . . .Chairman: Marvin H. Trattner. 848

"Construct Validity" . . . . Dr. Brian S. O'Leary. . . . 849

"Test of a New Model of Validity Generalization: Results for Tests Used in Clerical Selection" . . Kenneth Pearlman. 856
Frank L. Schmidt
John E. Hunter

"Synthetic Validity" . . . . Marvin H. Trattner. 879

SECTION 10. . . . . .STATISTICAL AND MEASUREMENT METHODOLOGIES. 883

"A Primer of Item Response Theory" . . . . . . Thomas A. Warm. 884

"A New Procedure to Make Maximum Use of Available Information When Correcting Correlations for Restriction in Range Due to Selection" . . . . Dr. James O. Boone. 906

"A Comparison of Three Models for Determining Test Fairness . . . . Dr. Mary A. Lewis 919

"A Method to Evaluate Performance Reliability of Individual Subjects" . . . . Alan E. Jennings. 933

"A Comparison of Two Criterion-Referenced Scoring Procedures for an Answer-Until-Correct, Multiple-Choice Performance Test" . . . . Dr. John B. Meredith, Jr. . 938
J. Thomas Martin, Jr.

"An Analysis of the OE Concept and Suggested Improvements" . . . . Dr. Clay E. George. 942
Henry L. Kinnison
H. Wayne Smith

SECTION 11. . . . . . .TESTING: TECHNIQUES AND TECHNOLOGIES. 948

"The Development of a Technique for Using Occupational Survey Data to Construct and Weight Computer-Derived Test Outlines for Air Force Specialty Knowledge Tests (SKTs)" . . . . . William J. Phalen 949

"Evaluation of Computer-Derived Test Outlines Using Conventional Test Outlines as a Criterion Reference During Test Development Projects" . . . . . CAPT Conrad G. Bills. . . . 976

"A Generalization of Sequential Analysis to Decision Making with Tailored Testing". . . . . Dr. Mark D. Reckase 994

"A Methodology to Evaluate the Aptitude Requirements of Air Force Jobs" . . . . . . Lloyd D. Burtch 1012

SECTION 12. . . . . .EXAMINATION ITEMS. 1026

"Objective Evaluation of Correspondence Course Items" . . . . Dr. Andrew N. Dow 1027

"The Emergence of an Item-Writing Technology" . . . . Gale Roid 1035
Tom Haladyna

SECTION 13. . . . . TRAINING PROGRAMS AND PROBLEMS. 1067

"Aircrew Training Research - Project ACTIVE" . . . . CAPT W. E. Keates 1068

"Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" . . . . R. A. Dapra 1091
W. Byham
M. G. Rumsey
A. Castelnovo
Dr. Richard S. Wellins

"How Do You Buy 'Good Design': An Examination of the Army's TEC Program". . . CAPT Robert R. Begland. . . 1098

"Content Validation of Class A School Curricula in the Coast Guard" . . . . Michael J. Bosshardt. . . . 1107
David A. Bownas
Richard S. Lanterman

"Experimental Evaluation of a High Technology Training Program". . . . . Dr. Arthur Kahn 1116

SECTION 14. . . . .INSTRUCTIONAL EVALUATION AND TEST DEVELOPMENT. 1137

"The Instructional Quality Inventory: Introduction and Overview". . . . . Dr. John A. Ellis 1138
Wallace H. Wulfeck II
Robert E. Richards
Norman D. Wood
M. David Merrill

"Design of Machine Scorable 'Hands On' Performance Tests in a Paper and Pencil Mode". . . . . Robert N. Johnson 1161

SECTION 15. . . . . .PERFORMANCE FEEDBACK 1180

"A Learning-Receptive State as Induced
by an Auditory Signal or Frequency
Pulse . . . . Dr. Raymond O. Waldkoetter. 1181
Dr. John R. Milligan

SECTION 16. . . . . .PERFORMANCE MEASUREMENT. 1192

"Complexity of Flight Path Data as an
Index of Skill in Piloting Performances
from a Flight Simulator Based Job-
Sample Test". . . . . Brian D. Shipley, Jr. . . . 1193

"Evaluation of Intelligence Producing
Capability of Selected Combat Arms
Units". Earl Rubright 1205
Alvaline B. Jackson

"Learning Aptitude, Error Tolerance, and
Achievement Level as Factors of
Performance in a Visual-Tracking Task". . . Brian D. Shipley, Jr. . . . 1220

SECTION 17. . . . . .SIMULATORS AND SIMULATION. 1248

"Evaluation of Troubleshooting Simulator". . . Dale A. Steffen 1249
Anita S. West

"Methodology for Evaluating Operator
Performance on Tactical Operational
Simulator/Trainers" . . . . Dr. Charles W. Howard . . . 1255

"Critical Performances of Battalion
Command Groups. . . . . Dr. Ira T. Kaplan 1264
Herbert F. Barber

"An Application of Tactical Engagement
Simulation for Unit Proficiency
Measurement". . . . . Dr. C. Mazie Knerr. 1316
Robert T. Root
LTC Larry E. Word

"Evaluation of the MODIA Planning
System" . . . . CAPT John R. Welsh, Jr. . . 1335

"Disassociated Utility of Moribund
Brains" . . . . CDR C. F. Meredith. 1373

REPORT OF STEERING COMMITTEE 1375

BY-LAWS OF THE MILITARY TESTING ASSOCIATION 1376

INDEX OF AUTHORS AND LIST OF CONFEREES 1383

MTA KEYNOTE ADDRESS:

QUALITY OF LIFE

RADM W. H. STEWART, USCG

Chief, Office of Personnel

Thank you Capt. Ferguson.

On behalf of the Coast Guard, I want to add my personal welcome to each of you to the twentieth annual MTA conference. This conference will cap two decades of effort to exchange technical information and know-how in the personnel management area. For twenty years many of you have made special efforts to present scholarly papers. For twenty years, each of the services has made special efforts to host this conference. That the conferences have continued for twenty years is a testimony to their worth. That your membership and attendance now includes representatives from the academic communities, from other government agencies, from private industry, and from military services of other countries is also an indication that much of the information you seek to exchange is of a broad and possibly universal nature.

At this time, I would like to welcome in particular Colonel Seuberlich and Dr. Puzicha from Germany and also Squadron Leader Thompson from Australia, as well as Mr. Beel from the Royal Navy, and Colonel Leach from Canada. I understand that Canada has volunteered to host this convention at Toronto in 1980. I also want to recognize Mr. Foley of the Navy Personnel Research and Development Center who will host this convention in San Diego next year. I wish to welcome the participants from the Universities, from private industry, and from other government agencies. Also, I want to acknowledge the presence of the commanding officers of our Coast Guard training units and their staffs who have been attending the Commanding Officer/Training Officer Conference.

I have reviewed the proceedings of your last three meetings and, though I am not scientifically qualified to judge the merits of your papers, I can say, as a qualified layman, that you generate a considerable amount of material. Considering the volume, complexity, and specificity of your output, I'm not certain whether I admire most the people who are delivering this information or the ones who are receiving and understanding it. In any case, it is not hard to understand why you always have a full schedule.

As Chief of Personnel of the Coast Guard, I am very concerned with utility, efficiency, and productivity; for these are the measures of individual and organizational performance. As Chief of Personnel, I also wonder if the Coast Guard men and women of today can handle the Coast Guard of tomorrow. Most of the instruments and procedures that you develop are designed to help answer questions of this sort and to improve the efficiency or productivity of the organization supporting your research. However, such improvements may or may not benefit the individuals who are being managed. Almost always, the concept of utility ignores the individuals in the organization because the utility is designed to benefit the organization. This, or course, is good for the organization and what is good for the organization generally returns benefits to the individuals in the organization. But, I have seen some great exceptions. We have kept ships at sea too long; and we have permitted long work days, and long work weeks. Of course, when there is a valid need which requires personal sacrifice, not many will complain. But we cannot justify working people 16 hour days simply because there is utility to it. The point is that we must be aware of the very real balance between benefits to the organization and to the individual. One way of benefiting both the individual and the organization is to increase professionalism at all personnel levels. This is the major goal of the Commandant of the Coast Guard. That is, we will encourage and assist professional development of benefit to the individual and of value to the organization.

It is the balance between individual and organizational benefits which determines the quality of life. This balance must continually be reestablished as conditions change. For example, the Coast Guard has just approved the policy to provide the opportunity for women to serve in all billets on board all ships and stations and in all grades and ranks, including the billet of commanding officer. The only restriction is that adequate personal privacy can be provided. This decision was not the product of an organizational utility model. For one thing we have always felt that the Coast Guard has done well even without women. Therefore, the decision to open opportunities for women was based primarily on considerations of equity and justice. This decision is the product of those social, philosophical, political, and legal forces which are continuously evolving and changing our society. So, with one value judgment, a huge change has been introduced into the Coast Guard.

One of the objects of the change was, to make greater opportunities available to women. But, as I indicated before, I am also concerned with both individual and organizational performance; and certainly there is no intention or putting women (or men, for that matter) into positions where they are not qualified or where they cannot perform adequately. Not only would such assignment be unfair to the individual woman (or man), but also it would reduce organizational performance levels. We all know that the ability to perform is a function of individual aptitude, training and motivation. If a woman wants to serve in a previously all male billet or job, then she has (by definition) adequate motivation to perform. But, she may not be qualified because of lack of training or experience even though she has the aptitude. This is also true for most male recruits we take into the Coast Guard.

The question has been, and still is, who can best be trained, that is, who has the aptitude for training? It is in this area that your classification tests have made a valuable contribution. But, do these tests work equally well for women as for men? Our mechanical aptitude tests are effective in predicting mechanical learning ability and knowledge for the white male majority; but most women, and many minorities, perform at the chance score level on these tests. This implies either, that most women and many minorities have no mechanical aptitude of use to the Coast Guard, or that we have not yet built tests that are culture-fair in evaluating their ability to learn to do mechanical work. Both my staff and I believe that the tests are culturally biased. Even so, how can I implement a policy which permits women to go into enlisted ratings which are heavily loaded with mechanical skill requirements, if all we know about women's mechanical aptitudes is that they score at the chance level on our mechanical tests. It is obvious to me that we need new test instruments (which we are now building) to tell us about the mechanical abilities of women and minorities.

Of course we also need to know if the tests are valid predictors of performance both in school and on the job.

One of the major problems has been a lack of knowledge about the job. However, I expect that your efforts in the job-task analysis will provide basic information which can be used to evaluate and validate not only the test instruments, but also the curriculums of our training schools, and even the structure and composition of the job itself. This effort is extremely important. It has been estimated that a work appraisal system for Civil Service could cost the entire Federal Government a half a billion dollars a year. However if such a system could increase productivity by as much as two percent it would effect savings far outweighing its cost.

These considerations of course involve technical questions which you as professionals in this field must answer with empirical studies. These studies, I am told, must conform to the new Uniform Guidelines on Employee Selection Procedures just released. I understand that the intent of the Uniform Guidelines is to assure equity and justice and to mandate fair recognition of the individual's potential, regardless of group membership. So, it seems that the social, philosophical, political, and legal forces have resulted in producing these guidelines; just as they did in our Coast Guard decision to provide equal opportunities to women.

However, these great and elegant decisions cannot be fully implemented without the supporting technologies to help the organization adapt to these changes.

We ask your help to assist us with your technology as we accommodate to t change that equity dictates. Change by itself, threatens organizational efficiency. This change (especially in personnel) is greater today than ever before. Change presents a problem. This has always been true. For example a young Naval officer wrote in his journal,

> Change thus succeeding change with bewildering rapidity...find all who have sought to keep up...have been called upon to absorb new ideas before the last has been assimilated.

This was written in 1879--almost a hundred years ago.

If change is handled properly, it can improve the quality of service life, maintain or improve productivity, and increase the level of professionalism.

I believe this is your mission. You are responsible for the research and development efforts needed to supply us with new tools, instruments, and procedures and knowledges which help us as managers to effectively accommodate to new situations. I am also confident that you will anticipate future changes, and even become instruments of change yourselves.

I am sure you will rise to this occasion because it is, after all, your life work. To the extent that you are always concerned for the individual, and assume a responsibility to improve the quality of life of each man and woman in the service, both the individual and the organization will benefit.

This is my belief, but only you can make it happen.

Thank you.

20th Annual Conference of the
MILITARY TESTING ASSOCIATION

Officers, Chairmen, and Committee

MTA President
CAPTAIN JAMES E. FERGUSON

MTA Secretary
MR. JOHN A. BURT

Chairmen:	Committee:
RICHARD C. WILLING	Program
CWO4 JOHN E. SCHWARTZ	Financial
CWO3 LARRY N. MONROE	Audiovisual
LCDR CLINTON W. CARTER	Social
LT LARRY C. YOUNG	Registration

20th Annual Conference
MILITARY TESTING ASSOCIATION

30 October - 3 November

MONDAY, OCTOBER 30

Lobby	
1200-1900	Registration
Presidential Suite (Room 404)	
1600-1800	Steering Committee Meeting
Gazebo Room	
1900-2000	Informal Reception

TUESDAY MORNING, OCTOBER 31

South Ballroom	
0900-1000	Conference called to order
0900-0915	Greetings by Coast Guard Institute Commanding Officer CAPT JAMES E. FERGUSON
0915-1000	Keynote Address RADM W. H. STEWART, USCG Chief, Office of Personnel
1000-1030	Break
1030-1140	INTERNATIONAL PRESENTATIONS

"Strain by Prolonged Duty Hours and Problems as to Mobility of Soldiers - As Seen by Federal Armed Forces Association" (20 min.)
COL H.E. SEUBERLICH, German Federal Armed Forces Association

"Execution of Large Occupational Analysis of the Royal Navy's Operations Branch" (20 min.)
C.D. BEEL, Royal Navy

"A Strategy for Task Analysis and Criterion Definition Based on Nonmetric Multidimensional Scaling" (30 min.)
LCOL GLENN M. RAMPTON, Canadian Forces Personnel, Applied Research Unit

1140-1200 Announcements

1200-1300 Lunch

TUESDAY AFTERNOON, OCTOBER 31

Appaloosa Room
1300-1430 PERSONNEL APPRAISAL

"Quality of ROTC Accessions to the Army Officer Corps" (15 min.)
DR. ARTHUR C.F. GILBERT and DR. RICHARD S. WELLINS, Army Research Institute, and DR. JOHN I. WELDON, U.S. Army Training and Doctrine Command

Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB)" (30 min.)
JOHN J. MATHEWS AND LONNIE D. VALENTINE, JR., Brooks Air Force Base, WAYNE S. SELLMAN, Randolph Air Force Base

Appaloosa Room
1500-1630 EXAMINATION ITEMS
Evaluating and Improving

"Objective Evaluation of Correspondence Course Items" (30 min.)
DR. ANDREW N. DOW, USNETPDC

"The Emergence of an Item-Writing Technology" (30 min.)
GALE ROID and TOM HALADYNA, Oregon State System of Higher Education

Arabian Room
1300-1430 METHODS OF DETERMINING PERSONNEL AVAILABILITY

"PAM: A Methodology for Predicting Air Force Personnel Availability" (20 min.)
H. ANTHONY BARAN, ANDREW J. CZUCHRY, JOHN C. GOCLOWSKI, DUNCAN L. DIETERLY, FREDRIC F. PHILLIPS, STUART E. PESKOE, and ANTHONY J. LOFASO, Air Force Human Resources Laboratory

Symposium: Methodology for Mobilization Population Inventory
Chairman: DR. JACK M. HICKS

"Some Implications of Commercial Test Normings for Mobilization Surveys"
R. F. BOLDT, Educational Testing Service

"Measuring the Military Base Population of the 1980's"
M. A. FISCHL, US Army Research Institute

20

"Development of a Mobilization Population Inventory Using Existing ASVAB Data Banks"
GEORGE V. RUX and WILLIAM GRAHAM, Military Enlistment Processing Command

"Air Force Experience with PROJECT TALENT"
LONNIE D. VALENTINE, JR., Air Force Human Resources Laboratory

Arabian Room
1500-1630

PERFORMANCE MEASUREMENT

"Complexity of Flight Path Data as an Index of Skill in Piloting Performances from a Flight Simulator Based Job-Sample Test" (25 min.)
BRIAN D. SHIPLEY, JR., US Army Research Institute Field Unit, Fort Rucker, Alabama

"Evaluation of Intelligence Producing Capability of Selected Combat Arms Units" (40 min.)
EARL W. RUBRIGHT, 80th MTC/NSA
ALVALINE JACKSON

"Learning Aptitude, Error Tolerance, and Achievement Level as Factors of Performance in a Visual-Tracking Task" (25 min.)
BRIAN D. SHIPLEY, JR., US Army Research Institute Field Unit, Fort Rucker, Alabama

Palomino Room
1300-1430

VALIDATION-PREDICTION, Session 1

"The Impact of Valid Selection Procedures on Workforce Productivity" (25 min.)
FRANK L. SCHMIDT, ROBERT C. McKENZIE, and TRESSIE W. MULDROW, U.S. Civil Service Commission and JOHN E. HUNTER, Michigan State University

"Job Performance of USAF Bypassed Specialists" (20 min.)
CAPT WILLIAM H. CUMMINGS and CAPT DAVID S. VAUGHAN, USAF Occupational Measurement Center

"Analysis of Heavy Equipment Operator Jobs" (25 min.)
SIDNEY A. FINE, HOWARD C. OLSON, DAVID D. MYERS, and MARGARETTE C. JENNINGS, Advanced Research Resources Organization

Palomino Room
1500-1630

VALIDATION-PREDICTION, Session 2

"Predictive Utility of the Officer Evaluation Battery (OEB)" (15 min.)
DR. ARTHUR C.F. GILBERT, US Army Research Institute

"Assessment Center Variables as Predictors of On-Job Performance Characteristics" (25 min.)
DR. CHARLES H. CORY, NPRDC

"Using an Assessment Center to Predict Leadership Course Performance of Army Officers and NCOs" (25 min.)
FREDERICK N. DYER and RICHARD E. HILLIGOSS, Army Research Institute Field Unit, Fort Benning, Georgia

"Validity of Associate Ratings of Performance Potential by Army Aviators" (15 min.)
ROBERT F. EASTMAN, US Army Research Institute Field Unit, Fort Rucker, Alabama, and MARIE LEGER, US Army Research Institute

WEDNESDAY MORNING, NOVEMBER 1

Appaloosa Room
0800-0935

OCCUPATIONAL-TASK ANALYSIS, Session 1
Issues and Answers

"Obstacles to and Incentives for Standardization of Task Analysis Procedures" (20 min.)
ROBERT W. STEPHENSON and HENDRICK W. RUCK, Air Force Human Resources Laboratory

"Task Analysis: Destination or Journey" (15 min.)
DR. MELVIN D. MONTEMERLO and DR. FRANK M. AVERSANO
US Army Training Support Center

"Four Fundamental Criteria for Describing the Tasks of an Occupational Specialty" (20 min.)
DR. WALTER E. DRISKILL and CAPT FRANK C. GENTNER, USAF Occupational Measurement Center

"Two Applications of Occupational Survey Data in Making Training Decisions" (20 min.)
CAPT DAVID S. VAUGHAN, ATC Technology Applications Center
CAPT JOHN R. WELSH

"The Stability Over Time of Air Force Enlisted Career Ladders as Observed in Occupational Survey Reports" (20 min.)
WALTER E. DRISKILL and FREDERICK B. BOWER, JR., USAF Occupational Measurement Center

Appaloosa Room
1100-1130

OCCUPATIONAL-TASK ANALYSIS, Session 2
Using Instructional Systems Development

"The Collection and Prediction of Training Emphasis Ratings for Curriculum Development" (20 min.)
HENDRICK W. RUCK, NANCY A. THOMPSON, AND SQDN LDR DAVID C. THOMSON, USAF Human Resources Laboratory

"Data Base to Determination of Training Content: A Manageable Solution" (20 min.)
D.D. DAVIS, CNET

"Using the Computer to Build the Task Inventory" (15 min.)
THOMAS M. ANSBRO, CNET

"Systematic Instructional Validation Through Testing" (35 min.)
DR. MARJORIE A. KUENZ and FREDERICK C. ROBERTS, III
Naval Health Sciences Education and Training Command

Arabian Room
0800-0930

STATISTICAL AND MEASUREMENT METHODOLOGIES, Session 1

"A Primer of Item Response Theory" (30 min.)
THOMAS A. WARM, US Coast Guard Institute

"A New Procedure to Make Maximum Use of Available Information When Correcting Correlations for Restriction in Range Due to Selection" (30 min.)
DR. JAMES O. BOONE, Civil Aeromedical Institute, Federal Aviation Agency

Arabian Room
1000-1130

STATISTICAL AND MEASUREMENT METHODOLOGIES, Session 2

"A Comparison of Three Models for Determining Test Fairness" (25 min.)
DR. MARY A. LEWIS, Civil Aeromedical Institute, Federal Aviation Agency

"A Method to Evaluate Performance Reliability of Individual Subjects" (15 min.)
ALAN E. JENNINGS, Civil Aeromedical Institute, Federal Aviation Agency

"A Comparison of Two Criterion-Referenced Scoring Procedures for an Answer-Until-Correct, Multiple-Choice Performance Test" (20 min.)
DR. JOHN B. MEREDITH, JR. and J. THOMAS MARTIN, JR., Data-Design Laboratories

"An Analysis of the OE Concept and Suggested Improvements" (30 min.)
DR. CLAY E. GEORGE and HENRY L. KINNISON, Texas Tech University and H. WAYNE SMITH

Palomino Room
0800-0930

VALIDATION-PREDICTION, Session 3

"Performance Test Objectivity: Comparison of Interrater Reliabilities of Three Observation Formats" (30 min.)
GERALD J. LAABS, Navy Personnel R & D Center
WILLIAM A. NUGENT

"Prediction of Field Artillery Officer Performance" (15 min.)
ARTHUR C.F. GILBERT, RAYMOND O. WALDKOETTER, and ANTHONY E. CASTELNOVO, US Army Research Institute

Palomino Room
1000-1130

VALIDATION-PREDICTION, Session 4
Symposium: Innovative Test Validation Strategies
Chairman: MARVIN H. TRATTNER

"Construct Validity"
BRIAN S. O'LEARY, U.S. Civil Service Commission

"Test of a New Model of Validity Generalization: Results for Tests Used in Clerical Selection"
KENNETH PEARLMAN and FRANK L. SCHMIDT, U.S. Civil Service Commission, JOHN E. HUNTER, Michigan State University

"Synthetic Validity"
MARVIN H. TRATTNER, U.S. Civil Service Commission

WEDNESDAY AFTERNOON, NOVEMBER 1

Appaloosa Room
1300-1430

OCCUPATIONAL TASK ANALYSIS, Session 3
Instructional Systems Development (ISD) and NEPDIS Overview

"Scheduling Formal School Training to Maximize Cost Effectiveness" (20 min.)
DOUG GOODGAME, Texas A&M University

"Methods for Determining Safety Training Priorities for Job Tasks (20 min.)
NANCY A. THOMPSON and HENDRICK W. RUCK, Air Force Human Resources Laboratory

Appaloosa Room
1500-1630

OCCUPATIONAL-TASK ANALYSIS, Session 4
Applying Task Analysis Methodology

"Methods for Collecting and Analyzing Task Analysis Data" (20 min.)
A. JOHN ESCHENBRENNER and PHILIP B. DeVRIES, McDonnell Douglas Astronautics Co., HENDRICK W. RUCK, Air Force Human Resources Laboratory

"Methodology for Selection and Training of Artillery Forward Observers Job Analysis" (20 min.)
JOHN B. MOCHARNUK and RUTH ANN MARCO, McDonnell Douglas Astronautics Co.

"Observer Self-Location Ability and Its Relationship to Cognitive Orientation Skills" (30 min.)
JOHN R. MILLIGAN and RAYMOND O. WALDKOETTER, Army Research Institute Field Unit, Fort Sill, Oklahoma

"Job Analysis in the US Army Medical Training Environment" (20 min.)
J. S. TARTELL, US Army

Arabian Room
1300-1430

SIMULATORS AND SIMULATION, Session 1
Design, Evaluation, and Personnel Performance

"Evaluation of Troubleshooting Simulator" (30 min.)
DALE A. STEFFEN and ANITA S. WEST, Denver Research Institute

25

"Methodology for Evaluating Operator Performance on Tactical Operational Simulator/Trainers" (30 min.)
DR. CHARLES W. HOWARD, Army Research Institute, Fort Bliss, Texas

"Critical Performances of Battalion Command Groups" (30 min.)
IRA T. KAPLAN and HERBERT F. BARBER, Army Research Institute, Fort Leavenworth, Kansas

Arabian Room
1500-1630

SIMULATORS AND SIMULATION, Session 2
Design, Evaluation and Personnel Performance

"An Application of Tactical Engagement Simulation for Unit Proficiency Measurement" (45 min.)
C. MAZIE KNERR and ROBERT T. ROOT, Army Research Institute, LTC LARRY E. WORD, US Army Training Support Center

"Evaluation of the MODIA Planning System" (45 min.)
CAPT JOHN R. WELSH, JR., Air Training Command, Lackland AFB, Texas

Palomino Room
1300-1430

PERSONNEL SELECTION

"Weighted Selection System for AFROTC Applicants--Perspective After Second Year of Use" (20 min.)
LT COL DAVID K. JACKSON and M. MERIWETHER GORDON, JR., AFROTC/ACME

"The Defense Language Aptitude Battery (DLAB)" (20 min.)
ROBERT G. HENDERSON, Defense Language Institute, Foreign Language Center

"Monte Carlo Computer Programs for Simulating Selection Decisions from Personnel Tests" (30 min.)
J.W. THAIN, Defense Language Institute, Foreign Language Center

Palomino Room
1500-1630

GENERAL

"Computer Assisted Reference Locator (CARL) System: An Overview" (25 min.)
WILLIAM A. SANDS, Navy Personnel Research and Development Center

THURSDAY MORNING, NOVEMBER 2

Appaloosa Room
0800-0930

OCCUPATIONAL-TASK ANALYSIS, Session 5
CODAP, Occupational Analysis for Training and Task Consolidation

"CODAP: A New Modular Approach to Occupational Analysis" (20 min.)
MICHAEL C. THEW and JOHNNY J. WEISSMULLER, Air Force Human Resources Laboratory

"Occupational Analysis for Field Grade Army Officers (30 min.)
SALLY J. VAN NOSTRAND, Army Research Institute, and M. REID WALLIS, Richard A. Gibboney Associates

"A Technique for Selecting Electronic Specialties for Consolidation" (20 min.)
HENDRICK W. RUCK, Air Force Human Resources Laboratory

Appaloosa Room
1000-1130

USING RATING SCALES
Issues, Evaluations, and Applications

"The Content Issue in Performance Appraisal Ratings" (35 min.)
CAPT R.H. MASSEY, C.J. MULLINS, and J.A. EARLES, Air Force Human Resources Laboratory

"Differential Responses on Alternately Anchored Job Rating Scales" (20 min.)
LT COL JIMMY L. MITCHELL, Air Force Occupational Measurement Center

"Sample Size and Stability of Task Analysis Inventory Response Scales" (20 min.)
JOHN J. PASS and D.W. ROBERTSON, Navy Personnel Research and Development Center

"Benchmark Scales for Collecting Task Training Factor Data" (15 min.)
SQN LDR DAVID C. THOMSON and KEN GOODY, Air Force Human Resources Laboratory

Arabian Room
0800-0930

OCCUPATIONAL SURVEYS, Session 1
Collecting, Evaluating, and Using the Data

"Civilian Ground Safety Officer Job and Training Requirements Survey" (15 min.)
DOUGLAS K. COWAN, Air Force Human Resources Laboratory

"Determining the Training Requirements of United States Coast Guard Warrant and Commissioned Officer Billets" (20 min.)
J.W. CUNNINGHAM, North Carolina State University
D.W. DREWES

"Evaluating the Army Occupational Survey Program Methodology: Answer Booklets, Questionnaire Length, and Population Coverage" (25 min.)
EUGENE M. BURNS, US Army Military Personnel Center

"The Use of Job Satisfaction Data in the Occupational Survey Program" (20 min.)
CAPT JOHN X. OLIVO and CAPT ELENA J. WEBER,
USAF Occupational Measurement Center

Arabian Room
1000-1130

OCCUPATIONAL SURVEYS, Session 2
Symposium: US Army Job Satisfaction and Retention Project

"General Overview and Initial Findings of the Project on Job Satisfaction and Retention of U.S. Army Enlisted Personnel"
LAWRENCE A. GOLDMAN, DARRELL A. WORSTINE, and CEDELLA J. BONETTE, US Army Military Personnel Center

Palomino Room
0800-0930

TRAINING PROGRAMS AND PROBLEMS, Session 1

"Aircrew Training Research - Project ACTIVE" (45 min.)
CAPT W.E. KEATES, Canadian Armed Forces

"Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" (20 min.)
R.A. DAPRA and W. BYHAM, Development Dimensions, Inc., M.G. RUMSEY, A. CASTELNOVO, and R.S. WELLINS, Army Research Institute

Palomino Room
1000-1130

TRAINING PROGRAMS AND PROBLEMS, Session 2

"How Do You Buy 'Good Design': An Examination of the Army's TEC Program" (25 min.)
CAPT ROBERT R. BEGLAND, TRADOC HDQTS.

"Content Validation of Class A School Curricula in the Coast Guard" (30 min.)
MICHAEL J. BOSSHARDT, DAVID A. BOWNAS, Personnel Decisions Research Institute, RICHARD S. LANTERMAN, U.S. Coast Guard

"Experimental Evaluation of a High Technology Training Program" (35 min.)
DR. ARTHUR KAHN, Westinghouse D&ES Center

THURSDAY AFTERNOON, NOVEMBER 2

Appaloosa Room
1300-1430

TESTING: Techniques and Technologies

"The Development of A Technique for Using Occupational Survey Data to Construct and Weight Computer-Derived Test Outlines for Air Force Specialty Knowledge Tests (SKTs)" (30 min.)
WILLIAM J. PHALEN, Air Force Human Resources Laboratory

"Evaluation of Computer-Derived Test Outlines Using Conventional Test Outlines as a Criterion Reference During Test Development Projects" (20 min.)
CAPT CONRAD G. BILLS, USAF Occupational Measurement Center

"A Generalization of Sequential Analysis to Decision Making with Tailored Testing" (20 min.)
MARK D. RECKASE, University of Missouri-Columbia

"A Methodology to Evaluate the Aptitude Requirements of Air Force Jobs" (20 min.)
LLOYD D. BURTCH, Air Force Human Resources Laboratory

Arabian Room
1300-1430

PERFORMANCE FEEDBACK

"A Learning-Receptive State as Induced by an Auditory Signal or Frequency Pulse" (30 min.)
DR. RAYMOND O. WALDKOETTER and DR. JOHN R. MILLIGAN, Army Research Institute Field Unit, Fort Sill, Oklahoma

Palomino Room
1300-1430

INSTRUCTIONAL EVALUATION AND TEST DEVELOPMENT

"The Instructional Quality Inventory: Introduction and Overview" (20 min.)
JOHN A. ELLIS, WALLACE H. WULFECK II, Navy Personnel Research and Development Center, ROBERT E. RICHARDS, NORMAN D. WOOD, The Pennsylvania State University, M. DAVID MERRILL, Courseware, Inc.

"Design of Machine Scorable 'Hands On' Performance Tests in a Paper and Pencil Mode" (60 min.)
ROBERT N. JOHNSON, US Army Administration Center

THURSDAY EVENING, NOVEMBER 2

Gazebo Room
1900-2000

Social Hour

Ballroom
2000-2200

Dinner

FRIDAY MORNING, NOVEMBER 3

South Ballroom
0900-1030

WOMEN IN THE UNIFORMED SERVICES

"Differential Field Assignment Patterns for Male and Female Soldiers" (20 min.)
"DR. L.W. OLIVER and MS. N.E. BABIN, Army Research Institute

"The Premature Attrition of Navy Female Enlistees" (20 min.)
GERRY L. WILCOVE, PATRICIA J. THOMAS, and CONSTANCE BLANKENSHIP, Navy Personnel Research and Development Center

"Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Description of Leader Behavior" (30 min.)
MAJ JEROME ADAMS, JACK M. HICKS, Army Research Institute

"Female Utilization in Non-Traditional Areas" (20 min.)
JOSEPH A. BERGMANN and RAYMOND E. CHRISTAL, Air Force Human Resources Laboratory

SECTION 1

OCCUPATIONAL SURVEYS

CIVILIAN GROUND SAFETY OFFICER JOB AND TRAINING REQUIREMENTS SURVEY

By
Douglas K. Cowan
Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper are those of the author and are not necessarily those of the United States Air Force.

I. INTRODUCTION

This study was the result of an expressed need by the Air Force Inspection and Safety Center (AFISC) to determine job types existing within the civilian ground safety officer area and to identify training requirements essential to the career development of the job incumbents. Consequently, the objectives of this study were to identify significant job types within the civilian ground safety officer population and the job characteristics which differentiate the identified job types from one another; to compare the task training emphasis recommended by job incumbents to the training emphasis placed on tasks within the Ground Safety Officer course (CIPO5D); and to construct a recommended career progression ladder for civilian ground safety officers comparable to that which exists for Air Force enlisted members, inasmuch as no career progression ladder currently exists for civilian ground safety officers.

II. METHOD

The job inventory used to collect job information from the civilian ground safety officers was developed by the Air Force Inspection and Safety Center (AFISC), with the assistance of the USAF Occupational Measurement Center (OMC) and the Air Force Human Resources Laboratory (AFHRL). The inventory was based upon Air Force job survey procedures spelled out in AFR 35-2, Occupational Analysis. It consisted of a background information section, which included personal and job-related data items, and a list of 295 significant work tasks organized under eleven major duty headings. In the background information section, each incumbent was questioned concerning formal education, pay grade, training courses completed, and other job-related items. The listing of tasks was reviewed by the incumbent for tasks performed in his current job. Each task performed was rated using a relative 9-point time spent scale to obtain an index that could be used to estimate how his time was distributed across all tasks in his job.

33

The job inventory was administered during April and May 1977 by AFISC to Department of the Air Force civilian employees who were assigned duty as ground safety officers and who had volunteered to complete the survey. A total of 212 job inventories was received from the field for analysis, which represented about fifty percent of the population.

An identical duty and task listing, but with a 9-point scale to reflect training emphasis recommended for each task, was sent to approximately 50 civilian ground safety officers at duty locations across the continental United States to obtain an estimate of needed training emphasis on each task. Forty-six civilian ground safety officers voluntarily completed the ratings and returned the survey booklets for analysis.

A similar 9-point rating scale using the same tasks and duties was forwarded to the School of Engineering, Arizona State University, Tempe, Arizona, to obtain training emphasis ratings from instructors in the Ground Safety Officer Course (CIPO5D)[1] to gain an estimate of current emphasis placed on training for the tasks listed in the job inventory.

III. RESULTS

Job Survey Analyses

Job analyses were performed using several of the Comprehensive Occupational Data Analysis Programs (CODAP) described by Archer (1977), Christal and Ward (1967), Morsh and Christal (1966), and Christal (1974). Six specific job types were identified through the hierarchical grouping process. Figure 1 shows the six job types and the grouping diagram. Nominal titles were assigned to each job type, based upon a functional analysis of the incumbents' job titles and assignment information.

Although six job types were identified through the grouping process, two of the groups, GRP 006 and GRP 028, appeared to be major command specific and, therefore, outside of a normal career progression route. Figure 2 depicts a conceptualized career ladder based upon an analysis of the hierarchical clustering of the sample and the level of the job as determined by background information supplied by incumbents. The civilian career ladder depicted is strikingly similar to the airman career ladder presented in AFR 39-1, Airman Classification Regulation, for the safety specialty, AFSC 241X0.

---

[1] Course offered by Arizona State University under Government contract.

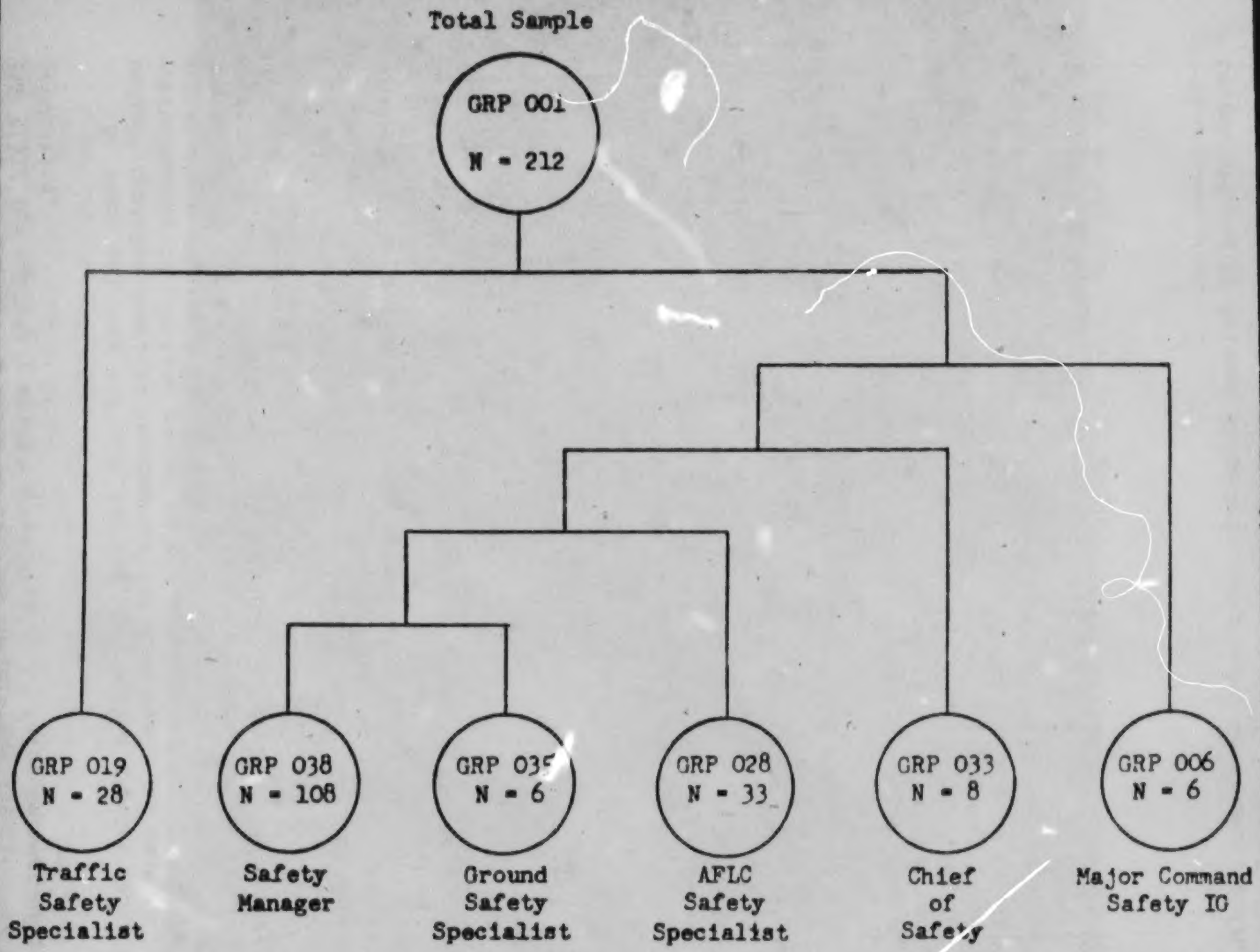
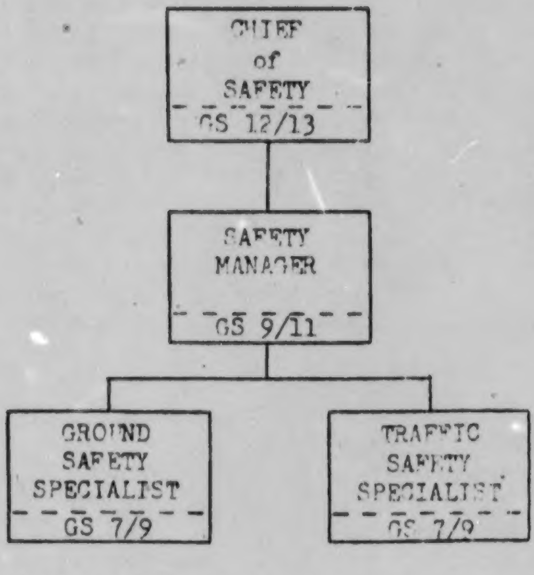

Total Sample
GRP 001
N = 212
GRP 019
N = 28
Traffic Safety Specialist
GRP 038
N = 108
Safety Manager
GRP 035
N = 6
Ground Safety Specialist
GRP 028
N = 33
AFLC Safety Specialist
GRP 033
N = 8
Chief of Safety
GRP 006
N = 6
Major Command Safety IG

Figure 1. Cluster Diagram of Ground Safety Officer Job Types


CHIEF of SAFETY
GS 12/13
SAFETY MANAGER
GS 9/11
GROUND SAFETY SPECIALIST
GS 7/9
TRAFFIC SAFETY SPECIALIST
GS 7/9

Note: Grades indicated are based on average grade levels reported

Figure 2. Conceptualized Career Progression Ladder Derived from Hierarchical Clustering

The average estimated percent time spent by the members of the six job types was summed by duty. The results are displayed in Table 1. The most time-consuming duty for each job types has been circled to illustrate the primary function of the group. The distinction between the ground safety specialist, the traffic safety specialist, and the AFLC safety specialist was rather clear-cut. However, the differences for the managerial job types were not as clearly evident. Both the safety managers and chiefs of safety spread their time across all duties, but the members of the chief of safety group spent more than 57% of their time in supervisory tasks (duties A, B, C, & J), while the safety managers spent only 38% of their time in the same duties. While the Major Command Safety Inspector General group and the Chief of Safety group are the most difficult to distinguish, it should be noted that the Major Command Safety Inspector General group spends more time in supervisory duties A, B, and C than the Chief of Safety group (52.18% vs. 44.59%), and more time in forms, records, and reports, duty E (9.14% vs. 5.06%); but spends less time in coordination, duty J (7.03% vs. 12.82%) and no time in traffic safety, duty H (0.00% vs. 1.03%). The training function of the traffic safety specialist is clearly evident, in that nearly one-half of this group's time is spent in training (duties D and H), while all the other groups report relatively small percentages of time in the safety training duty.

Table 2 shows the average number of tasks performed by job type and the number of tasks performed at the 25, 50, 75 and 100% time spent level. The groups are ranked by the average number of tasks performed. Since it may be inferred that the larger number of tasks performed is an indication of a more diversified job, then it is apparent that the safety manager has the most varied job, while the major command safety inspector general performs the most specialized job.

A number of background variables exhibited important differences among the six job types. As shown in Table 3, the total count or mean was computed for the following variables: sex, grade level, education level, months in job, and months at base of assignment.

The selection and assignment of personnel to fill civilian ground safety officer positions is primarily from the male sector of the population, with females accounting for only [illegible]% of the total sample. A total of seven females was identified in the sample, but only five are represented in the six job types. It was found that the average grade level of the females fell nearly three grade levels below the male, and a difference description computed between male and female members indicated that generally the female employees performed in a clerical rather than a technical or supervisory capacity.

Table 1. Estimated Percent Time Spent by Duty for Members of Each Job Type

Duty	Title	Percent Time Spent Ground Safety Spec.	 Traffic Safety Spec.	 Safety Manager	 Chief of Safety	 MAJ COM Safety IG	 AFLC Safety Spec.
A	Organizing and planning	7.06	7.63	11.24	16.27	18.97	9.21
B	Directing and implementing	7.34	11.74	10.29	13.01	16.76	8.93
C	Inspecting and evaluating	5.13	5.09	7.87	15.31	16.45	8.25
D	Training	1.93	8.94	3.86	4.68	3.09	2.43
E	Preparing & maintaining forms, records, and reports	7.35	8.77	6.66	5.06	9.14	7.48
F	Performing accident investigations	12.97	3.20	10.46	7.89	4.24	15.75
G	Performing site or facility safety inspections	18.18	2.83	18.25	15.29	13.33	32.33
H	Conducting traffic safety training and education	0.35	40.37	2.73	1.03	0.00	0.75
I	Preparing ground accident indices	31.34	3.47	15.15	3.73	5.24	2.98
J	Coordinating and maintaining liaison	5.33	5.36	8.89	12.82	7.03	7.87
K	Performing general unit safety functions	2.97	2.55	4.51	4.85	5.74	3.95
	Totals*	99.95	99.95	99.91	99.94	99.99	99.93

*Totals do not sum to 100% due to rounding error

Table 2. Average Number of Tasks Performed and Number of Tasks Performed by Selected Percentages of Time

Group	Average Number of Tasks Performed by Job Incumbents	Number of Tasks Accounting for Selected Percentages of Cumulative Time Spent on the Group Job Description 25%	50%	75%	100%
Safety Manager	188	41	92	155	294
Ground Safety Specialist	118	27	59	105	222
Chief of Safety	108	24	55	101	215
AFLC Safety Specialist	106	22	51	97	274
Traffic Safety Specialist	79	13	30	69	257
Major Command Safety IG	49	11	26	60	139

Table 3. Selected Background Variables by Job Type

Group	Sex Total Count M	F	Average GS Grade	Average Years of Education	Average Months in Job	Average Months on Base
Ground Safety Specialist	3	3	8.20	15.33	19	97
Traffic Safety Specialist	26	2	7.84	13.82	72	100
Safety Manager	108	0	10.98	14.42	54	81
Chief of Safety	8	0	12.14	16.38	54	154
Major Command Safety IG	6	0	12.00	14.83	34	100
AFLC Safety Specialist	33	0	9.91	14.56	49	113

The average grade by job type ranged from about GS-8 to GS-12, with the average grade of the total sample being slightly higher than GS-10. Members of all job types indicated rather high levels of education, with the chief of safety group members showing, overall, at least the attainment of a baccalaureate degree (or equivalent years of education) plus some additional education completed. The lowest average number of months in the job (19) was reported by the ground safety specialist, with the greatest number of months in the job (72) being reported by the traffic safety specialist. All groups reported fairly long base tenure.

Training Emphasis Analyses

Training emphasis analyses were completed using selected CODAP programs. Mean ratings of tasks provided by civilian job incumbents were computed. A Spearman rankorder correlation was computed between the mean recommended training emphasis ratings provided by the job incumbents and the percent of members performing the same tasks, resulting in an $r_s = .80$. A like correlation coefficient was computed for estimated percent time spent on the tasks with recommended training emphasis ratings, which produced an $r_s = .79$. Both correlation coefficents are significant at less than the .001 level of confidence, indicating that recommended training emphasis is very highly related to task performance. However, a substantial amount of variance (approximately 36%) in training emphasis is not accounted for by task performance alone and, as discussed by Ruck, Thompson, & Thomson (1978) in their paper, "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development," other factors such as consequences of inadequate performance, task delay tolerance, task difficulty, etc. must be considered. A Spearman rather than a Pearson correlation was computed, because neither percent of members performing nor percent time spent are normally distributed variables.

Table 4 shows the number of tasks in the total sample job description that received a mean training emphasis rating (2.53) or higher, the estimated percentage of time accounted for by these tasks, and the number of these tasks that were identified as being part of the Ground Safety Officer school curricula. Also shown is the total number of tasks identified in the job inventory as being taught in the school and the Spearman rankorder correlations of job incumbent training emphasis ratings with percent of members performing the tasks and estimated percent time spent on the tasks.

From Table 4 it can be seen that the Ground Safety Officer school provides training in less than half of the tasks with high recommended training emphasis (66 out of 140 tasks), but also provides training on 41 additional tasks which did not receive high recommended training emphasis.

Table 4. Relationship of Training Emphasis Ratings and Task Performance by Total Sample

Variable	Description	Number, Percentage or Correlation
A	Number of tasks with mean or higher recommended training emphasis ratings	140
B	Number of tasks with below the mean recommended training emphasis ratings	155
C	Percentage of job incumbent time accounted for by tasks in variable A	65%
D	Number of tasks in variable A included in Ground Safety Officer school	66
E	Total number of tasks identified as included in Ground Safety Officer school	107
F	Spearman rankorder corelation between recommended training emphasis ratings and percent members performing	.80
G	Spearman rankorder correlation between recommended training emphasis ratings and estimated percent time spent	.79

A Chi Square test was performed on tasks above and below the mean on recommended training emphasis versus whether the tasks were or were not being taught in the school. The computed Chi Square value was 12.74 (df = 1), which is significant beyond the .001 level of confidence. This finding indicates that the school put relatively more weight on teaching the tasks with higher, rather than lower, recommended training emphasis.

The percent time spent values for the taught and untaught tasks were summed separately for each duty (see Table 5). Inspection of the time spent values for taught and untaught tasks in the traditional management course-related duties (A, B, & D) and the nonmanagement duties (F, I, & K) revealed a much heavier emphasis by the school on the management areas than on the nonmanagement areas. The remaining unlisted duties contain a mixture of management, administrative, and worker-level tasks. The school emphasis on management is one reason why many of the tasks with higher recommended training emphasis were not being taught. Another reason is that some of these tasks are better taught by OJT.

Table 5. Estimated Time Spent on Taught and Untaught Tasks by Duty for Management Course-Related and Nonmanagement Duties

	Percent Time Spent	
Management Course-Related Duties	Taught	Untaught
A. Organizing and Planning	6.75	4.06
B. Directing and Implementing	6.47	3.84
C. Training	2.93	1.25
TOTAL	16.15	9.15
Nonmanagement Duties		
F. Performing Accident Investigations	2.97	7.19
I. Preparing Ground Accident Indices	4.07	7.40
K. Performing General Unit Safety Functions	.44	3.67
TOTAL	7.48	18.26

IV. CONCLUSIONS AND RECOMMENDATIONS

The use of the procedures established by AFR 35-2 in collecting job information from small populations appears to have produced high quality information similar to that attained from large military populations. Different job types were clearly identified through the use of CODAP, which allowed the conceptualization of a clear progression path for civilian ground safety officers. Since mean job incumbent recommended training emphasis ratings were very highly correlated to percent members performing and estimated percent time spent data, it must be assumed that these factors can be used interchangeably to account for most of the information contained in the training emphasis variable. It appears that a viable method for determining which tasks should receive training can be developed using the percent members performing and percent time spent data to determine at what career progression level tasks tend to be performed, and then the training emphasis data can be used to determine which tasks need special training. Summaries of background information provided valuable insight into the grade structure of the work force, as well as information about the educational level of the job incumbents and other pertinent information not readily available elsewhere.

The Ground Safety Officer course (CIP05D) appears to be fully supportive of the accident prevention program by providing management safety education to ground safety job incumbents, since 47% of the tasks that job incumbents rated fairly high on recommended training emphasis are also rated as being included in the Ground Safety Officer school. The remaining tasks receiving fairly high estimates of training emphasis appear to be tasks that could probably be trained during in-house training sessions, without recourse to formal school training.

From the conclusions, it appears that the following recommendations are in order:

1. That some form of career progression path similar to the one presented in this paper be established to formalize the present de-facto civilian career progression ladder.

2. That the relative priorities of technical and managerial skills and knowledges be determined by field interviews that would evaluate the consequences for job performance and career progression.

3. That consideration be given to assembling a panel of experts to "scrub down" the existing Ground Safety Officer course by systematically reviewing task training data on a task by task basis.

REFERENCES

AF Regulation 35-2, Occupational Analysis. Washington, D.C.: Department of the Air Force, 6 December 1976.

AF Regulation 39-1, Airman Classification Regulation. Washington, D.C.: Department of the Air Force, 1 June 1977.

Archer, W. B. Computation of group job descriptions from occupational survey data. PRL-TR-66-12, AD-653 654. Lackland Air Force Base, TX: Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Christal, R. E. The United States Air Force occupational research project. AFHRL-TR-73-75, AD-774 574. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Christal, R. E., & Ward, J. H., Jr. The MAXOF clustering model. In M. Lorr & S. B. Lyerly (Eds.), Proceedings of the conference on cluster analysis of multivariate data. New Orleans, LA: Catholic University of America, June 1967, 11.02-11.45.

Morsh, J. E., & Christal, R. E. Impact of the computer on job analysis in the United States Air Force. PRL-TR-66-19, AD-656 304. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, October 1966.

Ruck, H. W., Thompson, N. A., & Thomson, D. C. The collection and prediction of training emphasis ratings for curriculum development. Oklahoma City, OK: 20th Annual Conference of the Military Testing Association, 30 October-3 November 1978.

46

DETERMINING THE TRAINING REQUIREMENTS OF UNITED STATES COAST GUARD WARRANT AND COMMISSIONED OFFICER BILLETS

J. W. Cunningham and D. W. Drewes

North Carolina State University at Raleigh

Paper presented at the annual meeting of the
Military Testing Association
Oklahoma City, November 2, 1978

47

Problem and Purpose

Frequently changing duty assignments and staffing patterns in the U. S. Coast Guard create a continuing need for officers to acquire new knowledges and skills, which in many cases are best provided through formal training and education programs. Because of the high costs associated with such programs, however, it is essential that a systematic, empirical basis be established which will allow the Coast Guard to identify and provide within available funds the education and training most relevant to service requirements. It was in response to this need that the U. S. Department of Transportation contracted North Carolina State University to develop procedures and provide a data base that would allow the Coast Guard to assess its officer knowledge and skill requirements and to evaluate its postgraduate/post-commission education and training program against those requirements.

In designing a study for that purpose, we recognized that the military had historically used job/task analysis to establish job requirements, which, in turn, provided a basis for the development of training curricula. For lower-skill jobs employing large numbers of people, it is feasible to conduct such short-term training within military facilities. However, the small numbers of people involved and the level, types, and diversity of professional and technical knowledge required make it infeasible in most cases for the Coast Guard to conduct the advanced training needed by its officers. For that reason, the Coast Guard has generally used colleges, universities, and other institutions to upgrade knowledges and skills in its officer ranks. Within that context, we thought it reasonable to define training requirements in terms of educational courses and training modules, rather than attempting to derive such requirements through the delineation of specific job tasks. Indeed, this approach seemed necessitated by the fact that higher education organizes its curriculum offerings into units, or courses, not specifically oriented to military requirements (or, for that matter, to the specific requirements of civilian jobs). Even disregarding this constraint, we would still have faced, under the more traditional approach, the problem of accounting for the multitudinous tasks involved in all of the Coast Guard's officer billet codes.

With the aforementioned purpose and considerations in mind, we outlined four major goals, or phases, for the study:

1. Phase 1 of the study involved the development of a survey questionnaire to provide information concerning officer billet requirements and resources in relation to the Coast Guard's postgraduate/postcommission education and training program (hereafter referred to as the PGC program). This questionnaire was designed to obtain respondents' ratings of (a) their billets' requirements for specified PGC courses and (b) their own competencies in relation to the same courses. In addition, the questionnaire sought certain biographical information, as well as information pertaining to the respondents' attitudes and opinions about various aspects of the PGC program.

2. Phase 2 involved the collection of questionnaire responses from a large, representative sample of Coast Guard officers and warrant officers.

3. Phase 3 consisted of descriptive statistical analyses of the questionnaire response data.

4. And Phase 4 called for an initial comparative analysis of educational and training requirements versus human resources in the Coast Guard's officer billet codes. This analysis involved comparisons between the respondents' billet and self ratings on specified educational and training courses.

Instrument Development

The data-gathering instrument in this study was titled the "Survey of Officer Billet Educational Requirements" (or SOBER). This questionnaire was divided into four main sections.

Section I: Biographical Information

Section I, titled "Information About You," was designed to provide background information on such factors as the respondent's current grade level, authorized grade of billet, specialty area, previous training and education, present educational activities, and educational plans. The 30 response items in this section were selected based on their potential usefulness in organizing and understanding the data obtained in the remainder of the questionnaire.

Sections II and III: Educational Requirements and Proficiencies

Sections II and III of the SOBER were designed to obtain information on (a) billet educational and training requirements and (b) officer knowledges and skills in relation to those requirements. Section II instructed the respondents to rate the requirements of their particular billets for the knowledges and skills represented in 681 course descriptions; Section III asked them to rate their own proficiencies in terms of the same courses. These 681 courses were selected from an original pool of over 5400 that were identified as potentially relevant to the

Coast Guard PGC program areas. The selections were based on program managers' and representatives' estimates of the importance of the various courses to the billets associated with their program areas.

The course descriptions comprising the items for Sections II and III were prepared by consultants at the various program-offering institutions. These consultants were instructed to divide each course into its major knowledge units (or topics) and to write a brief descriptive statement of each unit's content. In composite, the knowledge-unit statements comprised the course description. Two examples of these course-description items are shown in Figure 1. The 681 course items were arranged under seven major subject-field designations which, in turn, were subdivided into a total of 25 more specific subject categories (see Figure 2).

— —

Insert Figure 1 here

— —

— —

Insert Figure 2 here

— —

The respondent used a seven-point level-of-knowledge-required scale to rate his billet on the course items, and a corresponding seven-point scale to rate his own levels of knowledge relative to the same courses (see Figure 3). As shown, there is a point-for-point correspondence

— —

Insert Figure 3 here

— —

between the two scales. Billet ratings on all 681 courses were performed first, followed by self ratings on the same courses.

Section IV: Opinions About the PGC Program

Section IV, the last part of the SOBER questionnaire, asked the respondents for their personal opinions concerning various aspects of the PGC program. The questions in this section dealt with such topics as the respondent's personal goals in relation to the PGC program, the adequacy of certain program characteristics, and the acceptability of some possible program alternatives. Seven-point scales were used with 42 of the 57 items comprising this section, while the remaining items used scales containing two to six points, depending upon the question. Figure 4 shows two examples of these scales.

Insert Figure 4 here

Procedures and Results

The SOBER questionnaire was mailed to over 5,600 Coast Guard officers and warrant officers. Each officer received a package containing (a) the questionnaire, (b) a set of answer sheets, and (c) a franked return envelope pre-addressed to the North Carolina State University Center for Occupational Education. The officers were assured anonymity in their responses. Of the questionnaire returns received by the cut-off date, a total of 2,866 (51 percent) contained usable data. The numbers and percentages of usable returns by grade level are shown in Table 1.

Insert Table 1 here

Billet Requirements

Descriptive statistical analyses were performed on the billet knowledge-requirement ratings within each of the seven grade levels. Table 2 shows some results for the 10 subject areas that were most frequently required. The cell entries represent the numbers and proportions of courses in each subject area that were required in each of the seven grade levels.

Insert Table 2 here

As you can see in the bottom line of this table, the number of required courses increases monotonically with grade level. For example, 18 courses were required by warrant officers, 38 by lieutenants, 71 by commanders, and 189 by admirals. Language skills and personnel/manpower/psychology were the two most generally required subjects across grade levels. The data for these two areas suggest that all ranks require a core of knowledges and skills in communication, management, and human relations. Billets in the higher grade levels appear to require an elaboration of these knowledges and skills, as evidenced by the increased number of such courses as a function of rank.

For the most part, courses required at one grade level are also required at the higher levels, suggesting a progression of knowledge requirements as a function of rank. It appears, moreover, that the additional course requirements at successively higher ranks represent not just an elaboration of the core subject areas, but also the introduction of new areas. Quite evident, for example, are requirements in the higher ranks for courses in business management, law, and political science/government--subject areas for which there is relatively little requirement in the lower grade levels. In more general terms, an examination of cumulative course requirements with increasing grade level shows course acquisitions in four additional subject areas between the warrant officer and ensign/lieutenant, junior grade levels, three additional areas between lieutenants and lieutenant commanders, one additional area between commanders and captains, and six additional areas between captains and admirals. In contrast to the five areas of course requirements for warrant officers, admirals reported course requirements in 19 different subject areas. The unique configuration of courses at the admiralty level is assumed to reflect the broad-based responsibility for decision-making in all areas of Coast Guard activities. This inference is supported by the fact that the admirals report course requirements in such areas as business management, accounting/finance, economics, political science/government, law, and operations research. These courses are decision-theoretic and can be argued to reflect the decision-making requirements inherent in their billets.

At this point, I should introduce a qualifying note in connection with these data. As you know, billet characteristics and requirements vary within grade levels, and this variation is likely to be quite substantial. When the billet ratings are averaged within grade levels, many of the specific billet requirements are masked. This would be particularly true of the more technologically specific requirements. Thus, it is important that requirements also be analyzed by specialty area, by officer billet code, and by billet. Coast Guard Headquarters is, in fact, currently carrying out such analyses under the direction of Messrs. Joseph Cowan and Richard Lanterman. To date, they have examined selected OBC's and performed a cluster analysis of over 2,000 billets. It is worth noting that in addition to identifying a number of OBC- and cluster-specific requirements, their analyses support our previous findings in regard to core course requirements; that is, the language-skill and management-related requirements appear to be general across billet clusters and specialty areas as well as grade levels.

Billet Requirements Compared with Incumbent Knowledges

As mentioned, the SOBER questionnaire respondents also rated their own levels of knowledge in relation to the 681 course items. These individual self ratings were averaged within each of some 522 officer billet codes, yielding a mean "knowledge-resource" vector for each OBC. Individual billet-requirement ratings were also averaged within the 522 billet categories, producing a "knowledge-requirement" vector for each OBC. For each OBC, the knowledge-resource and knowledge-requirement vectors were then compared by means of a "requirement-resource disparity index" (or RRDI). This index represented the average resource deficiency per course, for those cases where the requirement estimate exceeded the resource estimate; that is, the average deficiency among those courses for which deficiencies were found within the particular OBC (see Figure 5).

Insert Figure 5 here

In general, both the RRDI value and the absolute number of course deficiencies tended to increase with grade level; and, consistent with our earlier findings (see Table 2), the course deficiencies tended to concentrate in the areas of language skills, personnel/manpower/psychology, and business management. These results again point to the increasing importance of certain core knowledges and skills as a function of rank. As noted in connection with the billet-requirement ratings, however, the results must also be examined by specialty area, by officer billet code, and by billet. Such analyses are currently underway at Coast Guard Headquarters.

Although our efforts in this area were somewhat exploratory, we believe that the requirement-resource disparity approach should have some potential use in assessing the educational and training needs of billets and billet clusters and, possibly, in assigning individuals to PGC training slots.

Opinions Concerning the PGC Program

The final set of analyses in this study were performed on the officers' responses to the questions about their opinions concerning the PGC program. The results are summarized in Figure 6.

Insert Figure 6 here

Among those respondents who had received PGC training, 87 percent felt that it had effected a moderate to great increase in their general performance. The percentage expressing this opinion ranged from 82 percent for warrant officers to 100 percent for admirals. Thus, at all grade levels, there appears to be a considerable perceived benefit from PGC training.

Only 30 percent of the respondents thought that their billets required a graduate degree. As shown in Figure 6, the percentage indicating a graduate degree requirement increased monotonically with grade level, and ranged from 7 percent for warrant officers to 66 percent for admirals. Professional and managerial requirements were the most frequently indicated reasons why a graduate degree was necessary in a billet.

The officers' most important educational goals were to improve their technical specialty skills and managerial capabilities, while the least important personal considerations were qualifying for licensing and increasing employability in civilian life. Their most important personal reasons for seeking PGC training were to extend their general knowledge and to a lesser extent, to enhance their promotional opportunities; however, professional licensing and prestige were unimportant considerations.

Among the various PGC program changes rated by the respondents, the most acceptable were (a) systematic evaluation of the schools and courses in the program, (b) periodic reviews of billet training requirements, (c) greater use of training facilities within commuting distance of the officer's permanent duty station, and (d) increased emphasis on management training (a preference congruent with the results of the billet-requirement analyses). On the other hand, the least acceptable PGC program changes included (a) the development of a Coast Guard postgraduate school as an alternative to civilian academic institutions, (b) a shorter postgraduate program supplemented with off-duty training, and (c) the civilianization of billets requiring scarce or unusual technical skills.

It would thus appear that the officers see the present PGC program as a means of enhancing their personal growth, improving their performance potential, and facilitating their advancement within the service. Although they favor a greater program evaluation effort and possibly some changes in program emphasis and site location, they do not seem to be seeking drastic changes in program philosophy and practice.

Some Initial Conclusions

Several initial conclusions were drawn based on these preliminary analyses. These conclusions are presented as tentative recommendations and are meant to be suggestive of the potential policy implications of the data.

The first conclusion is that all officers should be adequately trained in the core knowledge areas. The results of this study indicate that six language-skill courses, eight courses in personnel/manpower/psychology, and one business management course are judged to be required at all grade levels. All ranks from ensign up, excluding warrant officers, were judged to require nine language-skill courses, 11 personnel/manpower/psychology courses, four law courses, one math course, and three military short courses. These common requirements would seem to have implications for both pre-commission and post-commission training activities.

A second possible conclusion is that training opportunities should be provided at each grade level. The progression of knowledge requirements across grade levels argues well for specific training content oriented to rank. At each grade level a set of courses can be identified

58

such that if a course is judged to be required at that level, it will also tend to be required at each succeedingly higher level. Under a rank-specific training approach, training for knowledges not used at a particular rank might be deferred until that time at which the knowledges become important.

Related to our second conclusion is a third conclusion that rank-specific training content should be supplemented with billet- or OBC-specific training content. As mentioned earlier, the characteristics and requirements of the billets within a particular grade level may show considerable variation around the means for that level. Accordingly, it becomes necessary to take into consideration the unique requirements imposed by individual billets, OBC's or billet clusters.

Our fourth conclusion is that the Coast Guard should consider increasing the incidents of training opportunities provided each officer. The progression in the kind and quantity of knowledge requirements across ranks has already been mentioned. Knowledge requirements apparently shift from more technically and specifically oriented requirements at the lower ranks to the more people- and policy-oriented knowledges at the higher ranks. In order for the PGC program to be responsive to changing demands, training content must shift as a function of this demand. Unless we assume that training given at any stage in career progression will generalize to subsequent stages and will provide for all future knowledge demands, changing demand structure would appear to require training at successive points in officers' careers to prepare them for subsequent changes in performance requirements.

59

Our last and most obvious conclusion is that continued use should be made of the data base obtained in this study as a means of developing strategies to improve the match between training requirements and resources. The data obtained in this study represent a rich source of information that can be used to make informed decisions about the nature and scope of the training requirements among Coast Guard warrant and commissioned officer billets. As noted, however, the analysis performed in this study were of necessity primarily descriptive and limited in scope. A number and variety of additional analyses are needed to provide insight into the knowledge-requirement structure of these billets. Several such analyses are presently being conducted at Coast Guard Headquarters, and others are planned.

List of Project Reports

Cunningham, J. W., Drewes, D. W., Ondrizek, L. E., & Steele, D. L. Biographical information from United States Coast Guard officers and warrant officers relevant to the Coast Guard's post-commission education and training program. Report No. 2 (under Contract No. DOT-CG-51375-A with the U. S. Coast Guard, Department of Transportation). Raleigh: Center for Occupational Education, North Carolina State University, 1977, 124 pp.

Cunningham, J. W., Drewes, D. W., Ondrizek, L. E., & Steele, D. L. Attitudes and opinions of United States Coast Guard officers and warrant officers regarding the Coast Guard's post-commission education and training program. Report No. 3, 1977, 39 pp.

Cunningham, J. W., Drewes, D. W., Ondrizek, L. E., & Steele, D. L. Rated post-commission educational and training requirements of officer and warrant officer billets in the United States Coast Guard. Report No. 4, 1977, 232 pp.

Cunningham, J. W., Drewes, D. W., Ondrizek, L. E., & Steele, D. L. An inventory of human resources among United States Coast Guard officers and warrant officers based on self ratings in knowledge areas related to the Coast Guard's post-commission education and training program. Report No. 5, 1977, 232 pp.

Cunningham, J. W., Drewes, D. W., & Steele, D. L. An analysis of educational and training requirements versus resources among officer and warrant officer billets in the United States Coast Guard. Report No. 6, 1977, 22 pp.

Cunningham, J. W., & Drewes, D. W. Determining the training requirements of United States Coast Guard warrant and commissioned officer billets: final report. Report No. 7, 1978, 113 pp.

Drewes, D. W., & Steele, D. L. Methods of selecting training. Report No. 1, 1976, 22 pp.

Figure 1. Examples of course-description items

422. ORGANIZATION AND MANAGEMENT

Introduction to Management--Role in modern society, the business organization as a system, management as a process, management in a changing environment.

Managerial Planning--Establishing objectives, formulating policy and operating plans, decision-making, organizational structure and relationships, delegation and decentralization, line and staff relationships, organization planning and change.

Social Aspects of Organizing--Organization as a social system, cultural background of organization, status systems, organization and the individual, staffing the organization.

Direction of the Organization--The employee as a person, leadership and motivation, communication, employee attitudes.

Controlling Organizational Performance--Basic factors in control, systems approach to managerial control, dysfunctional consequences of control, improving effectiveness of control, use of feedback in control.

571. FUNDAMENTALS OF WRITING

Review of English Grammar--Parts of speech, sentence structure, proper usage, punctuation.

Subject Matter of a Composition--Purpose, choosing and limiting a subject, selecting the major thesis, deciding what to way.

Organization--Basic principles of organization: making and refining the outline, introduction, ordering the parts of a composition, climax, conclusion.

Paragraphs--The paragraph as a single idea, paragraph organization and functions, topic sentences.

Writing Practice--Use of the fundamental principles of writing in composition of a variety of themes.

Figure 2. Outline of course-description items

ENGINEERING

- Bioengineering/Environmental Engineering
- Chemical Engineering
- Civil/Construction/Transportation Engineering
- Electrical/Electronics/Communications Engineering
- Industrial and Management Engineering
- Mechanical Engineering
- Metallurgical/Materials Engineering
- Naval Architecture/Marine Engineering/Ocean Engineering
- Engineering Mechanics
- Engineering Physics

MATHEMATICS/STATISTICS

- Mathematics
- Statistics

INFORMATION TECHNOLOGIES

- Computer and Information Sciences
- Operations Research

BUSINESS/MANAGEMENT/ADMINISTRATION

- Accounting/Finance
- Business Management
- Economics
- Personnel/Manpower/Psychology

PHYSICAL SCIENCES

- Physics
- Chemistry
- Other Physical Sciences

ARTS AND LETTERS

- Language Skills
- Literature/Philosophy
- History/Political Science/Government
- Law

INDUSTRY TRAINING PROGRAMS AND SELECTED SHORT COURSES

- Industry Training Programs and Selected Short Courses

Figure 3. Response scales used for the billet-requirement and self ratings

Level of Knowledge Required by the Billet

1 = No knowledge in this area is required by the billet.

2 = Little knowledge in this area is required by the billet.

3 = Some knowledge in this area is required by the billet.

4 = Moderate knowledge in this area is required by the billet.

5 = More than moderate knowledge in this area is required by the billet.

6 = Substantial knowledge in this area is required by the billet.

7 = Almost complete mastery in this area of knowledge is required by the billet.

Level-of-Knowledge Scale

1 = I have no knowledge in this area.

2 = I have little knowledge in this area.

3 = I have some knowledge in this area.

4 = I have moderate knowledge in this area.

5 = I have more than moderate knowledge in this area.

6 = I have substantial knowledge in this area.

7 = I have almost complete mastery in this area.

Figure 4. Examples of the response scales used with the attitude and opinion items

Section B

How important are (or were) the following reasons to you in desiring postgraduate/advanced training? Use the following scale:

Scale:

Blank = No opinion
1 = No importance
2 = Significantly below average importance
3 = Somewhat below average importance
4 = Average importance
5 = Somewhat above average importance
6 = Significantly above average importance
7 = Critical importance

Section C

How acceptable do you find the following alternatives to the present postgraduate/advanced training program? Use the following scale:

Scale:

Blank = No opinion
1 = Totally unacceptable
2 = Moderately unacceptable
3 = Slightly unacceptable
4 = Makes no difference
5 = Slightly acceptable
6 = Moderately acceptable
7 = Very acceptable

Figure 5. The requirement-resource disparity index (RRDI)

The knowledge-requirement and knowledge-resource vectors for each OBC number provided a basis for estimating the disparity between (a) the OBC's educational and training requirements and (b) the human resources in the OBC. This disparity estimate, termed the "Requirement-Resource Disparity Index" (RRDI), was computed for each OBC number as follows:

a. Each mean in the resource vector for a specified OBC number was subtracted from the corresponding mean in that OBC's requirement vector.

Requirement Vector	Resource Vector	Difference (d)
R_1	R_1'	$R_1 - R_1'$
R_2	R_2'	$R_2 - R_2'$
R_3	R_3'	$R_3 - R_3'$
.	.	.
.	.	.
.	.	.
R_{681}	R_{681}'	$R_{681} - R_{681}'$

b. All positive differences between means (+d) were retained; all negative differences between means (-d) were discarded.

c. The positive differences between the means in the two vectors were summed.

$$\Sigma(+d)$$

d. The RRDI value was obtained by dividing the sum of the positive differences by the number of positive differences.

$$\text{RRDI} = \frac{\Sigma(+d)}{k}$$

where $\Sigma(+d)$ = the sum of the positive differences and
k = the number of positive differences.

The resultant RRDI value represents the average difference between an OBC's requirement and resource estimates per knowledge (or course) item, for those cases where the requirement estimate exceeds the resource estimate. The k value, representing the number of items for which the requirement exceeds the resource, should also be of interest, as well as the $\Sigma(+d)$ value, representing the total estimated short-fall in knowledge resources. All three of these values should be considered in assessing the extent of the educational and training need for a particular OBC.

Figure 6. Summary of the respondents' opinions about the PGC program

1. What effect has PGC training had on your general performance?

   Moderate to great increase: 87%

2. Does the billet you rated require a graduate degree?

   Yes: 30% No: 70%

WO	Ensign/ Lt. JG	Lt.	Lt. Cmdr.	Cmdr.	Capt.	Admiral
7%	24%	30%	37%	50%	54%	66%

3. Personal educational goals?

   Most Important
   Improve technical specialty skills.
   Improve managerial skills.

   Least Important
   Develop competencies for licensing.
   Increase employability in civilian life.

4. Personal reasons for PGC training?

   Most Important
   Expand general knowledge base.
   Enhance promotional opportunities.

   Least Important
   Prepare for professional licensing.
   Increase social acceptance and prestige.

5. Acceptability of various changes in the PGC program?

   Most Acceptable
   Evaluation of PGC schools and courses.
   Periodic review of billet training requirements.
   Greater use of facilities within commuting distance.

   Least Acceptable
   Development of a Coast Guard PG school as an alternative to civilian institutions.
   Shorter PG programs.
   Civilianization of certain billets.

Table 1. Numbers and percentages of usable SOBER returns by grade level

Grade Level	Number	Percentage
1. Warrant Officers (WO-1, WO-2, WO-3, WO-4)	560	19.5
2. Ensigns and Lieutenants, Junior Grade	743	25.9
3. Lieutenants	670	23.4
4. Lieutenant Commanders	430	15.0
5. Commanders	295	10.3
6. Captains	153	5.3
7. Admirals	15	0.5
TOTAL	2866	99.9

68

TABLE 2. NUMBERS AND PROPORTIONS OF COURSES FROM 10 SELECTED SUBJECT AREAS THAT WERE REQUIRED IN EACH OF SEVEN GRADE LEVELS.

COURSE AREAS	WO	ENSIGN/ LT.JG	LT.	LT. CMDR.	CMDR.	CAPT.	ADMIRAL
1. INDUS. & MGT. ENGINEERING	0 .00	0 .00	0 .00	0 .00	1 .12	4 .50	4 .50
2. MATH	0 .00	2 .02	3 .05	2 .03	2 .03	4 .07	6 .10
3. ACCOUNTING/ FINANCE	0 .00	0 .00	0 .00	1 .08	3 .25	3 .25	11 .92
4. BUS. MGT.	1 .04	1 .04	4 .15	7 .26	13 .48	15 .56	23 .85
5. ECON.	0 .00	0 .00	0 .00	2 .06	3 .08	4 .11	17 .48
6. PERS./MAN-PWR/PSY.	8 .35	13 .56	11 .48	19 .83	21 .91	22 .96	23 1.00
7. LANG. SKILLS	6 .67	9 1.00	9 1.00	9 1.00	9 1.00	9 1.00	9 1.00
8. HISTORY/ POL. SCI./ GOV.	0 .00	1 .03	1 .03	2 .07	5 .17	14 .48	25 .86
9. LAW	0 .00	4 .12	5 .15	4 .12	8 .24	11 .32	27 .79
10. INDUST. TRNG. & SHORT COURSES	0 .00	5 .19	3 .12	5 .19	4 .15	8 .31	12 .46
ALL AREAS	18 .03	35 .05	38 .05	54 .08	71 .10	96 .14	189 .28

EVALUATING THE ARMY OCCUPATIONAL SURVEY PROGRAM METHODOLOGY: ANSWER BOOKLETS, QUESTIONNAIRE LENGTH, AND POPULATION COVERAGE

Eugene M. Burns

US Army Military Personnel Center
200 Stovall Street
Alexandria, Virginia 22332

THE VIEWS, OPINIONS AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY OR DECISION UNLESS SO DESIGNATED BY OTHER OFFICIAL DOCUMENTATION.

Evaluating the Army Occupational Survey Program Methodology: Answer Booklets, Questionnaire Length, and Population Coverage

Eugene M. Burns

US Army Military Personnel Center

In an on-going survey program, such as the Army Occupational Survey Program (AOSP), there exists the opportunity, as well as the need, to monitor and evaluate the survey methodology. Periodic evaluation efforts enable the survey managers to learn systematically from their survey experience and to improve and refine the survey procedures. On the basis of these evaluations, survey managers can modify their procedures to reduce cost, increase efficiency, or improve data quality. For example, the Bureau of the Census extensively evaluates its monthly Current Population Survey (U.S. Bureau of the Census, 1978). As an example of how an evaluation can be accomplished in a military survey program, this paper will discuss an experiment currently being conducted to evaluate various aspects of the AOSP survey methodology.

BACKGROUND

As of early 1978, the AOSP was programmed to survey about 100 Army enlisted Military Occupational Specialties (MOS) per year. The main portion of each survey was an MOS task inventory, but there were also sections covering background information, equipment, special requirements, and job satisfaction. The MOS with less than 1000 members (about two-thirds of the MOS) were surveyed in their entirety while the remaining MOS were sampled. (More detailed discussions of the AOSP are to be found in

(U.S. Department of the Army, 1977)). In early 1978, three aspects of the AOSP methodology seemed to be particularly in need of study:

1. The AOSP answer booklets. Prior to January 1978, AOSP (then known as MODB--The Military Occupational Data Bank) had used a single survey booklet. Responses were recorded in the booklet next to the questions. Starting in January 1978, separate question and answer booklets were introduced for economy reasons. The separate booklets were expected to be more difficult to use and, therefore, to yield less reliable data than the self-contained booklets, but the extent of this difference needed to be assessed.

2. Questionnaire length. Coinciding with the January 1978 answer booklet change, a 124 item job satisfaction section was added to the questionnaire.[1] Increasing the length of the questionnaire was also expected to have a deleterious effect on the quality of responses, especially towards the end of the questionnaire, where the job satisfaction section was located. Respondents might be too fatigued to give reliable responses to a section tacked on to the end of an already lengthy MOS questionnaire. Research was needed to determine whether the overall quality of responses to the questionnaire was affected by the addition of the job satisfaction section and, in particular, whether the job satisfaction section should be kept as part of the AOSP questionnaire.

---

[1] This section was copied from the November 1977 survey of Job and Career Satisfaction so that individual MOS could be analyzed against an Army-wide baseline.

3. Population coverage. Where sampling was required, AOSP surveys had relied on quota sampling. AOSP Project Officers at the installation level were mailed a number of questionnaires in proportion to the number of MOS incumbents assigned to their installation. The Project Officers were instructed to distribute the questionnaires to "personnel from as many different grades and duty positions as possible" (U.S. Department of the Army, n.d.: para 2-2). At issue was whether a shift to statistically more sound random sampling would be worth the effort involved in revamping the established distribution system, which was geared towards the operationally simpler quota sampling. The answer would depend, in large part, on a determination of the established system's effectiveness in attaining broad population coverage.

STUDY DESIGN

The experimental design shown in Figure 1 was proposed to investigate the effects of different answer booklets and of questionnaire length. By sending out the same questionnaire in two different formats (self-contained and separate answer booklets), the relative reliabilities of the two methods of recording answers could be determined. Similarly, by comparing questionnaires sent out with instructions to omit either the job satisfaction or the MOS-related sections with questionnaires which were fully completed, the effect on survey quality of the additional job satisfaction section could be estimated. Two types of comparisons to be made: (1) between individuals at the same point in time (e.g., between groups 1-2 and 3-4 at the first administration), and (2) within the same individuals at two different points in time (e.g., group 1 at the first and second administrations). The design in Figure 1 strengthens our ability to infer

Figure 1. Design for a Study of the Effects of Army Occupational Survey Program Answer Booklet Formats and Questionnaire Length on the Reliability of the Survey Data

Study Group	First Administration	Second Administration
1	Separate Answer Booklet	Separate Answer Booklet
2	Separate Answer Booklet	Self-contained Answer Booklet
3	Self-contained Answer Booklet	Self-contained Answer Booklet
4	Self-contained Answer Booklet	Separate Answer Booklet
5	MOS-related Only	MOS-related Only
6	Job Satisfaction Only	Job Satisfaction Only
7		Separate Answer Booklet
8		Self-contained Answer Booklet
9		MOS-related Only
10		Job Satisfaction Only

that observed differences are due to the experimental manipulation (e.g., answer booklet format) and not to other factors. Other factors could include (1) respondent familiarization with the questionnaire or resistance to a second questionnaire administration, and (2) changes in the work performed, reflecting either random monthly variation in tasks or increased soldier skill and responsibility. Groups 7-10 were included in the design to obtain estimates of the amount of change to be expected over the course of several months among soldiers who had not been exposed to the AOSP survey. (For further discussion of the logic of experimental design, see Campbell and Stanley, 1966).

EXECUTING THE STUDY DESIGN

Figure 1 describes a tightly controlled textbook experimental design. However, the design had to be embedded within an established survey program. Rather than gloss over the decisions and compromises entailed by this embedding, they will be described in detail in this section so that other survey programs may benefit from the AOSP experience.

How Many MOS? Questionnaires with separate answer booklets were being produced at the rate of roughly 10 a month, but any self-contained questionnaire would have to be produced by modifying an existing, separate answer booklet, questionnaire. Given the amount of effort involved in producing a high quality version of the standard booklet, it was decided to use just one MOS for the evaluation. Should the findings from one MOS prove ambiguous, the study could be expanded to more MOS. Sending several versions of more than one MOS survey might also unduly burden and confuse the AOSP Project Officers.

Which MOS? The decision to base the evaluation on an already existing questionnaire limited the MOS to one available in the spring of 1978. In addition, a large MOS was called for so that the evaluation would not interfere with the routine AOSP data requirements. The type of MOS chosen was not considered very important, although an MOS of paperwork specialists would not be suitable since these people would be expected to be more attuned to forms and complicated instructions than the typical soldier. Taking all criteria into consideration, the MOS which best suited the evaluation requirements turned out to be Motor Transport Operator (64C).

75

How to Sample? As stated above, the customary AOSP sampling procedure was quota sampling. It was necessary to decide whether the evaluation should rely on some more rigorous probability sampling scheme as called for in Figure 1's controlled experimental design, or whether it should also employ quota sampling. Since the experiment was designed to learn something about the operation of the on-going survey program, it was thought best not to make a major departure from the standard sampling procedures by insisting on a random selection of respondents at the first administration. If the 64C respondents were randomly selected from the 64C population, the 64C survey would be unique. Therefore, quota sampling was used to select first administration respondents. However, random selection of respondents would be absolutely necessary for the second administration. By randomly sampling persons who participated in the first administration, the analysis results could be generalized to that population. The second administration control groups were chosen after the first administration. Respondent distributions by sex, paygrade, and education were compared with the population distribution, sampling fractions were computed,and these fractions were used to randomly select additional soldiers for the second administration.

Sample Design. The method for obtaining respondents was chosen so as to place minimum strain on the AOSP distribution system. This could be accomplished by minimizing the number of installations to be affected by the study, which was done by choosing the eight installations with the largest 64C populations. At each of these installations, the regular AOSP quota was 11 percent, and an additional 11 percent were

76

chosen for the special conditions. Each installation chosen received all four versions of the 64C questionnaire (separate answer booklets, self-contained answer booklets, MOS-only, and job satisfaction-only). First administration questionnaires were distributed through the normal AOSP distribution channels. To achieve randomization of respondents among conditions, standard MOS-only, and job satisfaction-only booklets were intermixed in the shipping cartons. The self-contained version was shipped separately.

Given the use of quota sampling, there was no firm basis for determining the appropriate sample size needed for each experimental group. As a rough rule of thumb, sample size formula appropriate for random sampling was used to obtain a number which was then doubled to allow for attrition between administrations. At the 95 percent confidence level (for a normal probability distribution), the sample size was chosen to obtain a precision of ± 0.5 on the seven point scale used to gather task performance data. Using the equation

$$n = \frac{t^2 s^2}{d^2}$$

with s estimated as 2.0, the sample size obtained was 64 for each of the 10 study groups.

Questionnaire administrations were planned four months apart. The four month lag was decided upon after debriefing some soldiers after the administration of an earlier survey.

RESULTS

The evaluation described in this paper is still in progress. The most serious problem encountered thus far has been in-house personnel turbulence which delayed the shipment of the second administration questionnaire by nearly three months. As a result, retest plans for the instruction booklets (MOS only, job satisfaction-only) were dropped. One installation was unable to meet its suspense date on the first administration and was dropped from the study. Otherwise, only minor problems were encountered on the first administration. Table 1 presents the first administration return rates by installation and booklet type.

The analysis of the returns so far has focused on the representativeness of the sample by comparing the distribution of returns (all booklet types) with the 64C population distribution.

It must be noted that no such comparison can prove that questionnaire respondents were randomly sampled. Random sampling is a process, not a result which can be determined by post-hoc measurement. However, the more the respondent distribution approximates the population distribution, the easier it becomes to argue that the sampling procedure is producing results which are representative of the population.

The first question asked was whether the respondents were distributed among pay grades proportionate to the 64C population pay grade distribution. Of the seven installations, four departed significantly (at the .05 level) from the distribution expected on the basis of proportionate random sampling, as shown in Table 2. These four installations included some of the most conscientious and reliable AOSP Project Officers. Rather than reflecting unfavorably upon the AOSP Project Officers' conduct of their jobs, these departures from the expected distribution should be viewed as stemming from lack of explicit guidance calling for proportionate sampling. Summing over

Table 1. Return Rates by Installation and Booklet Type,

First 64C Administration

Installation	Standard		Self-Contained		Special Instructions	
	Sent	Accepted	Sent	Accepted	Sent	Accepted
Fort A	43	32	20	18	20	20
Fort B	101	77	50	38	50	36
Fort C	52	52	26	25	26	26
Fort D	149	148	74	70	74	66
Fort E	73	67	36	35	36	69[a]
Fort F	51	51	24	23	24	23
Fort G	57	50	28	28	28	29[a]
Total	526	477	258	237	258	271
		(90.7%)		(91.9%)		(105.0%)

[a] Some "standard booklet" soldiers were accidently given instructions to skip parts of the questionnaire, thus additional booklets were provided.

Table 2. A Comparison of the Actual 64C Respondent Distribution with the Expected Distribution, by Skill Level and Installation

	Fort A		Fort B		Fort C		Fort D		Fort E		Fort F		Fort G		All Installations	
Skill Level	Actual	Expd[a]	Actual[b]	Expd[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]
Skill Level 1																
E1-E2	29	15.87	24	24.32	12	29.26	19	24.85	42	28.38	27	14.39	17	17.72	170	158.36
E3	4	9.58	33	31.85	23	19.20	57	58.58	34	32.54	23	19.95	25	27.15	199	196.56
E4	19	27.18	85	85.58	35	35.79	134	125.37	51	73.87	33	37.22	43	39.12	400	423.27
Skill Level 2																
E5	17	12.09	26	24.32	20	12.86	54	50.37	31	24.46	11	17.27	16	14.73	175	156.09
Skill Level 3																
E6	1	3.93	3	5.84	9	3.91	20	23.74	10	7.83	3	5.56	6	5.98	52	55.95
Skill Level 4																
E7	1	2.35	6	5.09	5	2.98	8	9.09	3	3.92	2	4.61	0	2.30	25	30.77
Total	71	71.00	177	177.00	104	104.00	292	292.00	171	171.00	99	99.00	107	107.00	1021	1021.00
Chi square[c]	21.53		1.71		22.91		3.00		16.37		16.93		2.99		5.82	

[a] Expected frequency based on proportionate sampling of the installation 64C population.

[b] Excludes 23 anonymous respondents.

[c] With 5 degrees of freedom, $p_{.9} = 9.24$, $p_{.95} = 11.07$, and $p_{.99} = 15.09$.

the seven installations included in the study, we see that individual installation departures cancel each other out, so that the overall respondent distribution is not significantly different from the expected distribution. We may speculate that this result is not anomalous, and that overall AOSP samples generally lack consistent bias in coverage.

The second major question asked involved the distribution of respondents within pay grades E1 to E4. These pay grades collectively comprise skill level one under the new Enlisted Personnel Management System and include 76 percent of the 64C population at the seven installations. Within skill level one, there are three significant social groups: male high school graduates, male non-high school graduates, and females (virtually all of whom are high school graduates). Table 3 presents the results of a comparison of the actual respondent distribution with the expected distribution for the seven installations. In contrast with the preceding comparison, only two of the seven installations were found to depart significantly from the distribution expected on the basis of proportionate random sampling. These results are consistent with the hypothesis that, in general, Project Officers select respondents without regard for sex or educational background.

Taken as a whole, the findings of the representativeness study indicate that, while overall AOSP respondent distribution may be representative of the MOS, installation level distributions exhibit biases in the selection of respondents. If installation level results were ever desired, these biases would require weighting by pay grade to produce accurate results.

Table 3. A Comparison of the Actual 64C Respondent Distribution for Skill Level One (E1-E4) with the Expected Distribution, Sex, Civilian Education, and Installation

Sex and Education Group	Fort A		Fort B		Fort C		Fort D		Fort E		Fort F		Fort G		All Installations	
	Actual	Expd[a]	Actual[b]	Expd[a]	Actual	Expd[a]	Actual	Expj[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]	Actual	Expd[a]
Males																
High School Grads	38	36.32	119	105.70	38	43.83	139	132.56	78	78.83	51	48.15	48	49.60	511	498.14
Non-HS Grads	14	12.88	16	28.55	8	13.94	58	65.83	33	32.73	20	24.40	35	32.84	184	206.47
Females	0	2.80	7	7.75	24	12.23	13	11.61	16	15.44	11	9.45	2	2.56	73	63.39
Total, All Groups	52	52.00	142	142.00	70	70.00	210	210.00	127	127.00	82	82.00	85	85.00	768	768.00
Chi square[c]	2.98		7.26		14.63		1.41		0.03		1.22		0.32		4.23	

[a] Expected frequency based on proportionate sampling of the installation 64C Skill Level One population.

[b] Excludes 19 anonymous respondents.

[c] With 2 degrees of freedom, $p_{.9} = 4.61$, $p_{.95} = 5.99$, and $p_{.99} = 9.21$.

Plans are being formulated to extend these analyses to MOS to be surveyed during 1979 and to incorporate some of these quality control measures into the survey program. By studying installation sampling patterns over several surveys, it should be possible to determine where corrective measures such as providing feedback and/or additional guidance to project officers should be applied.

SUMMARY

The representativeness study was able to disclose patterns within the 64C respondent returns which were not apparent in the day-to-day operation of the AOSP. It is anticipated that the answer booklets and questionnaire length studies will similarly highlight aspects of AOSP methodology which might have gone unnoticed, or poorly noticed, without special effort at evaluation. An important result of the study has been the decision to incorporate some of these quality control measures into the survey procedures as a continual (rather than one-shot) methods evaluation. With each survey completed, on-going survey programs receive many opportunities to learn how to improve themselves. Statistical self-evaluations, such as those outlined in this paper, can be a valuable tool in taking advantage of those opportunities to learn systematically from experience.

REFERENCES

Campbell, Donald T. and Julian Stanley
    1966 Experimental and Quasi-Experimental Design for Research. Chicago: Rand McNally.

United States. Bureau of the Census
    1978 The Current Population Survey: Design and Methodology.

United States. Department of the Army
    1977 The Army Occupational Survey Program. Army Regulation 611-3.

    n.d. Army Occupational Survey Program Questionnaire Administration (Enlisted MOS). Department of the Army Pamphlet 611-3 (Draft).

THE USE OF JOB SATISFACTION DATA IN THE OCCUPATIONAL SURVEY PROGRAM[1]

John X. Olivo, Captain, USAF
and
Elena J. Weber, Captain, USAF

USAF OCCUPATIONAL MEASUREMENT CENTER
OCCUPATIONAL SURVEY BRANCH
LACKLAND AFB TX 78236

A paper presented at the Military Testing Association Convention
30 October - 3 November 1978

---

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Departement of Defense.

THE USE OF JOB SATISFACTION DATA IN THE OCCUPATIONAL SURVEY PROGRAM

John X. Olivo, Captain, USAF
and
Elena J. Weber, Captain, USAF

USAF Occupational Measurement Center
Occupational Survey Branch
Lackland AFB TX 78236

Each year the USAF Occupational Measurement Center conducts occupational analyses of 51 USAF airmen career ladders. The career ladders analyzed during any calender year vary from flight engineer to still photographer to dental technician. The data from these various career ladders are collected using a survey instrument which is divided into three parts: 1) specific biographical information about the survey respondent; 2) questions concerning the individual's job; and 3) a detailed listing of tasks. This paper will deal with the job satisfaction data collected in part two of the survey instrument. The four indices used to collect the job satisfaction data will be discussed first, followed by a brief review of the procedures used to compile the 1977 data. Next, uses of the data and trends noted from the 1977 data will be discussed. Finally, some applications of the data both within occupational surveys and also in training and management areas will be reviewed.

Four indices are used in a USAF job inventory to collect data concerning job satisfaction. The first is perceived job interest. Here the respondent is asked to rate how interesting he or she perceives his or her job on a seven point scale ranging from Extremely Dull to Extremely Interesting. The next two indices are perceived utilization of talents and training. A seven point scale which ranges from Not At All to Perfectly is used for these two indices. The final index of job satisfaction on the inventory is reenlistment intentions. Here the respondent is asked if he or she plans to reenlist. A four point scale ranging from No to Uncertain to Yes is used for this question.

This is the third year in which job satisfaction data has been compiled and used for comparison purposes with on-going surveys. Each year the format used to report the data has been changed. The 1975 data on survey respondents were combined with no divisions by time-in-service or career area group. The 1976 survey data were separated into two time-in-service groups, 1 to 48 months total active federal military service (TAFMS) and 49 plus months TAFMS. However, the 1977 data were sorted both by time-in-service and by career area groups. The three time-in-service groups used in 1977 summary statistics were 1-48 months TAFMS, 49-96 months TAFMS, and 97+ months TAFMS. This appeared to give the user sufficient distinction between the various time-in-service groups.

The problem of grouping the various Air Force specialties into career area groups was more difficult to resolve. An authoritative source document on which to base the groupings was necessary. It was decided to use AFM 26-3, Air Force Manpower Standards, (Vols II-V) as a basis for grouping the various career fields. The 67 enlisted specialties used for the 1977 summary were divided into seven groups. These were: Aircrew; Mission Equipment Operations; Mission Equipment Maintenance; Command Support; Medical; and Special Duty Identifiers. The list of the various Air Force Specialties comprising each of the seven groups is attached at the end of this paper.

The data are presented in a series of tables. Tables 1-3 present composite pictures of each of the three TAFMS groups by career area. This allows for easy identification of differences in each of the four job satisfaction indices from career area to career area for each of the three time-in-service groups.

The job satisfaction data presented in these tables has routinely been included as part of the occupational survey report (OSR). Although analyses of the data or plausible explanations for the data are not part of the report. The data are also presented for each of the job groups identified within the career ladder or career field being surveyed for time-in-service groups. Results from a particular field are then compared to the USAF average for the previous year to see if any large deviations exist. Large variations are highlighted in occupational survey reports.

In previous years the data had been arranged so that little direct comparison could be made. Having arranged the 1977 job satisfaction data to reflect time-in-service and career area groups has allowed more direct comparisons be made between current and previous surveys. For example, personnel with 49 to 96 months TAFMS in the administration career ladder, a specialty in the direct support career area, can be compared directly to other personnel with the same time-in-service from the direct support career area surveyed the previous year.

Several interesting trends were noted within the 1977 data. It had been assumed that when the data were organized by career area groups there would be some variance in each of the indicies from career area to career area. The assumption had been that clerical administrative personnel would not find their job as satisfying as would the dental technicians. The data, however, showed that across the career area groups the level of job satisfaction, perceived utilization of talents and training, and reenlistment intentions were fairly consistent. The major differences that occurred were between time-in-service groups, not career field groups. There typically was a slight (less than five percent) increase in job satisfaction from the 1 to 48 months TAFMS respondents to the 49 to 96 months TAFMS respondents. However, the increase between the 49 to 96 months TAFMS respondents and those with 97+ months TAFMS was fairly large, generally about ten percent. Again, the implications of these differences are not discussed in the OSR. Force managers, however, might and do find such data invaluable, and the Occupational Measurement Center is always ready to assist in interpreting and using these data.

There also appeared to be little connection between reenlistment intentions and the other three job satisfaction indicies. For survey respondents with 1 to 48 months TAFMS approximately three-fourths of the respondents in each career area group found their job interesting and felt their talents and training were being used fairly well or better; yet, less than half (46 percent) planned to reenlist. A good example were operating room personnel (AFS 902X2). While 80 percent or more of the first enlistment personnel found their job interesting and felt their talents and training were being used fairly well or better, only 35 percent planned to reenlist. This trend continued with the second term groups. Only among personnel with 97+ months TAFMS were the responses to the four indicies fairly consistent.

Another trend noted was that across all career area groups the level of job satisfaction was fairly consistent except for aircrew personnel. The level of job satisfaction among these personnel in each of the three time-in-service groups was well above that reported by incumbents in any other career area group. Unlike other career area groups, however, the aircrew personnel showed little, if any, increase in job satisfaction from one time-in-service group to the next. The only index that did increase markedly was the reenlistment intention.

Currently there are several agencies which use the job satisfaction data collected in occupational surveys. The Air Force Human Resources Laboratory at Brooks AFB, TX has continually used this data for a number of research projects. Headquarters Air Training Command at Randolph AFB, TX is attempting to develop some correlation between job satisfaction data and reenlistment rates to determine training effectiveness. Within the occupational survey program this data is primarily collected and reported for each individual specialty being surveyed. Occupational analysts sometimes find job groups within specialties which have consistently different ratings on the job satisfaction indices than other career ladder job groups. This might serve as another indicator for indentifying job type groups. In addition, analysts also report differences for particular specialty when compared to the other specialties within that career area group.

The job satisfaction data offers several areas for further research. One area would be to compare job satisfaction data among each year group within the 1-48 TAFMS group. Along this same line, personnel with 192 to 240 months TAFMS (the 16 to 20 year group) could be grouped individually and then compared to personnel with 97 to 191 months TAFMS. A second area of consideration would be a statistical analysis to determine whether in fact there are significant differences in job satisfaction data among the various career areas. Also, the relationship between Airmen Qualification Examination (AQE) scores and job satisfaction data should be further explored; if a relationship does exist, it would provide another piece of information that would help understand the complex work motivation issue.

Summary

The job satisfaction data collected from surveys conducted in 1977 were reported for time-in-service and career area groups. These data are routinely reported as part of the occupational survey report. While no detailed examination of the data is made, large deviations from other groups within the study or from the averages of the previous year are reported. These large deviations can sometimes be an aid in job typing. One consistent result is a low relationship between reenlistment intentions and the other three job satisfaction indicies. In addition to OMC, the job satisfaction data is used by HQ/ATC, AFHRL, and force managers at AFMPC and the Air Staff. Finally this data provides areas for future research into such issues as changing patterns in job satisfaction among year groups in the first four years of an air force career, determining the level of significance in job satisfaction among the various career areas, and the relationship between AQE scores and job satisfaction.

TABLE 1

EXPRESSION OF JOB INTEREST, PERCEIVED UTILIZATION OF TALENTS AND TRAINING AND REENLISTMENT INTENTIONS BY PERSONNEL WITH 1-48 MONTHS TAFMS SURVEYED DURING 1977*

	TOTAL SAMPLE	AIRCREW	MISSION EQUIPMENT OPERATIONS	MISSION EQUIPMENT MAINTENANCE	COMMAND SUPPORT	DIRECT SUPPORT	MEDICAL
I FIND MY JOB:							
DULL	16	3	25	17	12	14	15
SO-SO	19	6	25	21	15	14	15
INTERESTING	65	91	50	62	73	72	70
MY JOB UTILIZES MY TALENTS:							
NOT AT ALL OR VERY LITTLE	31	14	44	32	25	28	30
FAIRLY WELL TO VERY WELL	63	76	53	64	64	63	62
EXCELLENTLY OR PERFECTLY	6	10	3	4	11	9	8
MY JOB UTILIZES MY TRAINING:							
NOT AT ALL OR VERY LITTLE	26	14	26	26	20	25	17
FAIRLY WELL TO VERY WELL	66	64	67	67	67	64	69
EXCELLENTLY OR PERFECTLY	8	22	7	7	13	11	14
DO YOU PLAN TO REENLIST:							
NO OR PROBABLY NO	59	44	51	61	57	58	62
YES OR PROBABLY YES	41	56	49	39	43	42	48

* TO OBTAIN A REPRESENTATIVE SAMPLE, THE COMMAND SUPPORT AND MEDICAL AREAS CONTAIN RESPONSES COLLECTED DURING 1976 AND 1977

TABLE 2

EXPRESSION OF JOB INTEREST, PERCEIVED UTILIZATION OF TALENTS AND TRAINING AND REENLISTMENT INTENTIONS BY PERSONNEL WITH 49-96 MONTHS TAFMS SURVEYED DURING 1977*

	TOTAL SAMPLE	AIRCREW	MISSION EQUIPMENT OPERATIONS	MISSION EQUIPMENT MAINTENANCE	COMMAND SUPPORT	DIRECT SUPPORT	MEDICAL
I FIND MY JOB:							
DULL	13	3	27	12	11	16	14
SO-SO	16	8	19	16	15	16	11
INTERESTING	71	89	54	72	74	68	75
MY JOB UTILIZES MY TALENTS:							
NOT AT ALL OR VERY LITTLE	23	14	38	21	19	28	23
FAIRLY WELL TO VERY WELL	68	70	57	71	70	62	66
EXCELLENTLY OR PERFECTLY	9	16	5	8	11	10	11
MY JOB UTILIZES MY TRAINING:							
NOT AT ALL OR VERY LITTLE	24	11	28	22	18	28	18
FAIRLY WELL TO VERY WELL	66	63	64	68	71	63	67
EXCELLENTLY OR PERFECTLY	10	26	8	10	11	9	15
DO YOU PLAN TO REENLIST:							
NO OR PROBABLY NO	35	24	25	35	39	34	32
YES OR PROBABLY YES	65	76	75	65	61	66	68

* TO OBTAIN A REPRESENTATIVE SAMPLE, THE COMMAND SUPPORT AND MEDICAL AREAS CONTAIN RESPONSES COLLECTED DURING 1976 AND 1977

TABLE 3

EXPRESSION OF JOB INTEREST, PERCEIVED UTILIZATION OF TALENTS AND TRAINING AND REENLISTMENT INTENTIONS BY PERSONNEL WITH 97+ MONTHS TAFMS SURVEYED DURING 1977*

	TOTAL SAMPLE	AIRCREW	MISSION EQUIPMENT OPERATIONS	MISSION EQUIPMENT MAINTENANCE	COMMAND SUPPORT	DIRECT SUPPORT	MEDICAL
I FIND MY JOB:							
DULL	9	4	14	9	10	10	8
SO-SO	10	7	13	11	8	10	9
INTERESTING	81	89	73	80	82	80	83
MY JOB UTILIZES MY TALENTS:							
NOT AT ALL OR VERY LITTLE	15	8	23	14	16	17	12
FAIRLY WELL TO VERY WELL	65	65	64	68	57	62	66
EXCELLENTLY OR PERFECTLY	20	27	13	18	27	21	22
MY JOB UTILIZES MY TRAINING:							
NOT AT ALL OR VERY LITTLE	19	8	25	18	18	22	12
FAIRLY WELL TO VERY WELL	61	62	60	63	57	60	63
EXCELLENTLY OR PERFECTLY	20	30	15	19	25	18	25
DO YOU PLAN TO REENLIST:							
NO OR PROBABLY NO	27	20	31	28	27	27	23
YES OR PROBABLY YES	73	80	69	72	73	73	77

* TO OBTAIN A REPRESENTATIVE SAMPLE, THE COMMAND SUPPORT AND MEDICAL AREAS CONTAIN RESPONSES COLLECTED DURING 1976 AND 1977

LISTING OF MAJOR GROUPING AFSs

AIRCREW

1.	111X0	Defense Aerial Gunner
2.	112X0	In-Flight Refueling Operator
3.	113X0 A/C	Flight Engineer
4.	114X0	Aircraft Loadmaster
5.	115X0	Pararescue Recovery

MISSION EQUIPMENT OPERATIONS

1.	20XXX	Intelligence
2.	27XXX	Command Control Systems Operations
3.	29XXX	Communications Operations

MISSION EQUIPMENT MAINTENANCE

1.	30XXX	Communications Electronics Systems
2.	31XXX	Missile Electronic Maintenance
3.	32XXX	Avionics Systems
4.	34XXX	Training Devices
5.	36XXX	Wire Communications Systems Maintenance
6.	40XXX	Intricate Equipment Maintenance
7.	42XXX	Aircraft Systems Maintenance
8.	43XXX	Aircraft Maintenance
9.	44XXX	Missile Maintenance
10.	46XXX	Munitions and Weapons Maintenance

COMMAND SUPPORT

1.	10XXX	First Sergeant
2.	24XXX	Safety
3.	65XXX	Procurement
4.	66XXX	Logistics Plans
5.	67XXX	Accounting and Finance
6.	69XXX	Management Analysis
7.	70XXX	Administration
8.	71XXX	Printing
9.	73XXX	Personnel
10.	74XXX	Morale, Welfare, and Recreation
11.	79XXX	Information
12.	87XXX	Band

LISTING OF MAJOR GROUPING AFSCs (CONT)

DIRECT SUPPORT

1.	22XXX	Photomapping
2.	23XXX	Audiovisual
3.	25XXX	Weather
4.	39XXX	Maintenance Management Systems
5.	47XXX	Vehicle Maintenance
6.	51XXX	Computer Systems
7.	54XXX	Mechanical/Electrical
8.	55XXX	Structural/Pavements
9.	56XXX	Sanitation
10.	57XXX	Fire Protection
11.	59XXX	Marine
12.	60XXX	Transportation
13.	61XXX	Supply Services
14.	62XXX	Food Services
15.	63XXX	Fuels
16.	64XXX	Supply
17.	75XXX	Education and Training
18.	81XXX	Security Police
19.	82XXX	Office of Special Investigations and Counterintelligence
20.	92XXX	Aircrew Protection

MEDICAL

1.	90XXX	Medical
2.	91XXX	Medical
3.	98XXX	Dental

SPECIAL DUTY IDENTIFIERS (SDIs)

1.	99500	Recruiter
2.	99501	Engineering or Scientific Assistant
3.	99502	Military Training Instructor
4.	99503	United States Air Force Honor Guard
5.	99504	LGM-30 Facility Manager
6.	99505	Courier
7.	99506	Combat Information Monitor
8.	99508	Scatter Communications Maintenance Technician
9.	99509	Data Formatting Equipment Operator
10.	99600	Student Training Advisor
11.	99601	ICBM Maintenance Manager
12.	99602	Sensor Operator
13.	99603	Minuteman NCO Code Controller
14.	99604	Postal Specialist

October 1978

GENERAL OVERVIEW AND INITIAL FINDINGS OF THE PROJECT ON JOB SATISFACTION AND RETENTION OF U.S. ARMY ENLISTED PERSONNEL

LAWRENCE A. GOLDMAN, PH.D.
DARRELL A. WORSTINE
CEDELLA J. BONETTE

DISCLAIMER NOTICE

THE VIEWS, OPINIONS AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY OR DECISION, UNLESS SO DESIGNATED BY OTHER OFFICIAL DOCUMENTATION.

MILITARY OCCUPATIONAL DEVELOPMENT DIVISION
PERSONNEL MANAGEMENT SYSTEMS DIRECTORATE
U.S. ARMY MILITARY PERSONNEL CENTER
ALEXANDRIA, VA 22332

General Overview and Initial Findings of the Project on Job Satisfaction and Retention of U.S. Army Enlisted Personnel

I. Overview of the Job Satisfaction and Retention Project

The US Army Military Personnel Center's (MILPERCEN) job satisfaction and retention project was described at the 19th annual Conference of the Military Testing Association held in San Antonio, Texas in October 1977. The primary intent of today's presentation is to update the status of this project during the past year as well as to recapitulate its scope and what has been achieved up to now. This overview will consist of the following: (1) the environment in which the project was initiated; (2) project phases; and (3) the intended uses of the data.

A. Context of the Project:

Since 1968, MILPERCEN, through its Army Occupational Survey Program (AOSP) has systematically conducted occupational analysis of enlisted Military Occupational Specialties (MOS). In the fall of 1974, a job satisfaction section was added to each of its Duty MOS survey questionnaires. This section consisted of nineteen measures used to operationally define and empirically measure satisfaction with one's Army job and with military life. The definitions used essentially comprised the hygiene factors (intrinsic to one's job) and the motivator factors (extrinsic to one's job, relating to one's work environment) that Frederick Herzberg identified in his research on job satisfaction (Herzberg, Mausner, and Snyderman, 1959).

THESE NINETEEN FACTORS, AS SHOWN IN TABLE 1, PROVIDED VERY INCOMPLETE COVERAGE OF THOSE FACTORS WHICH COULD POTENTIALLY HAVE A SIGNIFICANT INFLUENCE ON JOB AND CAREER SATISFACTION. MOREOVER, THESE ORIGINAL FACTORS DID NOT PERTAIN DIRECTLY TO REENLISTMENT INTENT. CONSEQUENTLY, THE JOB SATISFACTION PORTION WAS EXPANDED TO MORE THOROUGHLY EXAMINE THE RELATIONSHIP OF JOB SATISFACTION (WORK ATTITUDES) TO THE RETENTION (DECISION TO STAY OR LEAVE THE SERVICE), UNIT MORALE AND DUTY PERFORMANCE OF ENLISTED PERSONNEL. THIS WAS BASED PRIMARILY ON RESEARCH CONDUCTED AT THE AIR FORCE HUMAN RESOURCES LABORATORY (ALLEY AND GOULD, 1975). INTEREST CENTERED ON THE RELATIONSHIP BETWEEN JOB AND CAREER SATISFACTION AND FIRST-TERM REENLISTMENTS.

THIS EXPANSION, CONSTITUTING THE INITIAL PHASE OF THIS PROJECT, WAS PART OF THE ARMY'S OVERALL EFFORT TO GAIN ADDITIONAL INSIGHTS INTO RETENTION, JOB SATISFACTION AND THE ALL-VOLUNTEER ARMY. THE PRIMARY GOAL IS TO IMPROVE THE ARMY'S ABILITY TO RECRUIT AND RETAIN AN ADEQUATE NUMBER OF QUALITY SOLDIERS.

AS TUTTLE AND HAZEL HAVE NOTED, THE MAJORITY OF THE RESEARCH AND APPLICATIONS CONCERNING JOB SATISFACTION HAVE OCCURRED IN INDUSTRY (TUTTLE AND HAZEL, 1974). WITHIN THE PAST TEN YEARS, HOWEVER, THE MILITARY HAS BEGUN TO APPLY RESEARCH FINDINGS FROM THE PRIVATE SECTOR AND TO SPONSOR ITS OWN RESEARCH IN THIS AREA. MILPERCEN'S EFFORTS IN THIS AREA RELATED TO THE QUALITY OF LIFE RESEARCH CONDUCTED BY THE AIR FORCE HUMAN RESOURCES LABORATORY

AND THE NAVAL PERSONNEL RESEARCH LABORATORY. AS PREVIOUSLY INDICATED, THIS JOB AND CAREER SATISFACTION PROJECT MOST CLOSELY RESEMBLES THE AIR FORCE'S APPROACH (ALLEY AND GOULD, 1975).

B. PHASES OF THE PROJECT

OUR PROJECT, WHICH HAS BEEN PROGRAMMED TO CONTINUE INTO FISCAL YEAR 1979, HAS CONSISTED OF THE FOLLOWING FOUR INTER-RELATED PHASES (THREE ARMY-WIDE ATTITUDE SURVEYS AND ONE OCCUPATIONAL SURVEY):

(1) ANALYSIS OF AN ARMY-WIDE SAMPLE SURVEY OF APPROXIMATELY 3800 SOLDIERS CONDUCTED IN AUGUST 1976. THIS EFFORT WAS CONFINED TO FIRST-TERM PERSONNEL IN PAYGRADES E-3 AND E-4 WHO PROVIDED INFORMATION ON 38 ITEMS CONCERNING JOB SATISFACTION AND REENLISTMENT INTENT. THE RESULTS, WHICH WILL BE SUMMARIZED LATER IN THIS PRESENTATION, WERE PUBLISHED IN MAY 1977. THE REPORT IS ENTITLED "JOB SATISFACTION AND REENLISTMENT INTENT FOR FIRST-TERM PERSONNEL - INITIAL FINDINGS".

(2) ANALYSIS OF RESPONSES OF APPROXIMATELY 4000 SOLDIERS ARMY-WIDE IN FEBRUARY 1977 TO AN 88 ITEM QUESTIONNAIRE. THIS EFFORT CONSTITUTED AN ABBREVIATED VERSION OF THE AIR FORCES' OCCUPATIONAL ATTITUDE INVENTORY, ALTHOUGH MODIFIED TO ACCOUNT FOR DIFFERENCES BETWEEN THE ARMY AND THE AIR FORCE. RESULTS OF THIS STUDY, WHICH WILL ALSO BE DESCRIBED IN THIS PRESENTATION, ARE TO BE PUBLISHED IN THE FIRST QUARTER OF FISCAL YEAR 1979. THE ANALYSIS ADDRESSES BOTH FIRST-TERM AND CAREER SOLDIERS AND COVERS: (A) THE MOST AND LEAST SATISFYING ASPECTS OF ARMY LIFE

AND WORK; (B) THE BEST PREDICTORS OF JOB SATISFACTION, REENLISTMENT INTENT, AND UNIT MORALE; (C) THE MOST IMPORTANT REASONS FOR ENLISTMENT AND SEPARATION; AND (D) THE RELATIONSHIP BETWEEN REENLISTMENT INTENT AND REENLISTMENT DECISION.

(3) ANALYSIS OF OCCUPATIONAL SURVEY DATA COLLECTED FROM MOS 00E (RECRUITER) AND MOS 79D (CAREER COUNSELOR) IN THE SPRING OF 1977. THIS PROJECT RELATES THE PERCEPTIONS OF 1100 RECRUITERS AND 150 CAREER COUNSELORS TO THOSE OF FIRST-TERM SOLDIERS ON MATTERS ASSOCIATED WITH ENLISTMENT AND SEPARATION. THIS REPORT IS ALSO SCHEDULED FOR PUBLICATION IN THE FIRST QUARTER OF FISCAL YEAR 1979.

(4) AN ARMY-WIDE SURVEY CONDUCTED IN NOVEMBER 1977 OF APPROXIMATELY 11,000 FIRST-TERM AND CAREER FORCE MEN AND WOMEN. THIS 362 ITEM QUESTIONNAIRE, REPRESENTING THE END PRODUCT OF OVER ONE YEAR OF DEVELOPMENTAL WORK, ADDRESSES THE ISSUES OF JOB SATISFACTION, REENLISTMENT INTENT, UNIT MORALE AND RECRUITER ACCURACY. IT ALSO COVERS THE IMPORTANCE OF FACTORS RELATED TO ENLISTMENT, SEPARATION OR RETIREMENT, AND REENLISTMENT. ANALYSIS HAS COMMENCED RECENTLY. INITIAL RESULTS (COVERING THE IMPORTANCE TO ENLISTMENT, REENLISTMENT AND SEPARATION OF THE FIRST-TERM FORCE) ARE TO BE PUBLISHED DURING THIS QUARTER. SUBSEQUENT ANALYSES WILL BE PUBLISHED INCREMENTALLY THROUGHOUT FISCAL YEAR 1979.

C. INTENDED USES

THE TWO PRINCIPAL USES OF THE ORIGINAL JOB SATISFACTION SECTION CONTAINED IN THE ARMY OCCUPATIONAL SURVEY PROGRAM

QUESTIONNAIRES, BASED ON A HERZBERG - BASED APPROACH, WERE: (1) TO DETERMINE THE DEGREE OF SATISFACTION/DISSATISFACTION BETWEEN AND WITHIN DIFFERENT OCCUPATIONAL SPECIALTIES, PARTICULARLY IF THESE MOS WERE IDENTIFIED AS "PROBLEM" MOS (DUE TO FACTORS SUCH AS A LARGE IMBALANCE BETWEEN AUTHORIZED AND OPERATING FORCE STRENGTHS, IMBALANCE BETWEEN CONUS AND OVERSEAS AUTHORIZATION; A LARGE NUMBER OF PERSONNEL EXPRESSING DISSATISFACTION WITH THEIR JOB, INTENDING TO SEPARATE OR RETIRE, AND/OR SPENDING A MAJORITY OF THEIR TIME ON NON-DUTY RELATED WORK); AND (2) TO AMPLIFY OTHER DATA COLLECTED IN THE QUESTIONNAIRE, INCLUSIVE OF DUTY/TASK INFORMATION AND SPECIAL KNOWLEDGES AND REQUIREMENTS.

RESULTS OBTAINED FROM THIS EXPANDED JOB SATISFACTION AND RETENTION PROJECT ARE INTENDED PRIMARILY TO MEET THE NEEDS OF KEY ARMY DECISION - MAKING AGENCIES (E.G., THE OFFICE OF THE DEPUTY CHIEF OF STAFF FOR PERSONNEL - RECRUITMENT AND REENLISTMENT DIVISION, AND THE ENLISTED PROMOTIONS AND SEPARATION BRANCH OF THE ENLISTED DIVISION) AS WELL AS THOSE OF CAREER COUNSELORS (REENLISTMENT NCOs THROUGHOUT THE ARMY).

IT WAS ALSO INTENDED THAT THIS PROJECT BE LINKED TO RELATED RESEARCH CONDUCTED BY OTHER ARMY AGENCIES AND OTHER SERVICES WITHIN DOD.

TO ACCOMPLISH THESE OBJECTIVES, CONSIDERABLE TIME WAS DEVOTED TO ASSESSING THE NATURE AND EXTENT OF OTHER COMPLETED STUDIES OR THOSE IN PROGRESS WITHIN DOD PERTAINING TO JOB SATISFACTION AND REENLISTMENT.

102

THE OUTCOME OF THIS ASSESSMENT WAS THE FOLLOWING LIST OF USES FOR THE DATA ANALYZED IN THIS PROJECT:

EXAMINATION OF RELATIONSHIPS BETWEEN JOB SATISFACTION AND:

- RETENTION (PARTICULARLY OF FIRST-TERM PERSONNEL)
- UNIT MORALE
- OCCUPATIONAL MISMATCH
- EFFECTIVE USE OF TRAINED ASSETS
- SELECTED STUDIES (E.G., WOMEN IN THE ARMY)

II. THE AUGUST 1976 ARMY-WIDE SURVEY

A. INTRODUCTION.

THE FIRST PHASE OF THIS JOB SATISFACTION AND RETENTION PROJECT CONSISTED OF ANALYSIS OF A SURVEY DISTRIBUTED TO A RANDOM SAMPLE OF PERSONNEL ARMY-WIDE IN AUGUST 1976. ALTHOUGH THIS QUESTIONNAIRE CONTAINED 80 ITEMS, ONLY 38 WERE ANALYZED, INCLUDING THE 17 INDEPENDENT AND TWO DEPENDENT FACTORS (OVERALL JOB SATISFACTION AND REENLISTMENT INTENT) USED IN THE JOB SATISFACTION PORTION OF THE AOSP QUESTIONNAIRES FOR WHICH DATA HAVE BEEN COLLECTED SINCE 1974. THE OTHER 19 FACTORS IN THE QUESTIONNAIRE USED IN THIS ANALYSIS WERE THOSE INSERTED BY OTHER ARMY AGENCIES FOR THEIR OWN SPECIFIC PURPOSES. IT SHOULD BE NOTED THAT ALL 80 ITEMS WERE CAST IN FINAL FORM PRIOR TO THE INITIATION OF THIS PROJECT. SINCE THIS QUESTIONNAIRE WAS A COMPOSITE REPRESENTING THE NEEDS OF DIFFERENT AGENCIES, IT WAS THEREFORE NOT DESIGNED TO BE A "COMPREHENSIVE" INSTRUMENT FOR MEASURING THE PRIMARY FACTORS INFLUENCING THESE TWO CRITERION MEASURES. AS PREVIOUSLY STATED, COVERAGE OF FACTORS WITH THE POTENTIAL OF

MEASURING REENLISTMENT INTENT WAS MINIMAL THROUGH USE OF THE 19 FACTORS USED IN THE AOSP. WITH THE ADDITION OF THESE 19 OTHER FACTORS, COVERAGE OF FACTORS THAT COULD MEASURE REENLISTMENT BEHAVIOR WAS SUBSTANTIALLY IMPROVED BUT NOT COMPLETE. IN SUBSEQUENT SURVEYS (E.G., THE FEBRUARY 1977 ARMY-WIDE SURVEY WHICH WILL BE DISCUSSED LATER IN THIS PRESENTATION), THE MAJOR DEFECTS IN THE COVERAGE OF REENLISTMENT RELATED FACTORS, AND TO A LESSER EXTENT IN THE COVERAGE OF JOB SATISFACTION RELATED MEASURES, HAVE BEEN REDUCED CONSIDERABLY. THE ANALYSIS OF THE AUGUST 1976 SURVEY WAS BASED ON 3,679 PERSONNEL IN PAYGRADES E-3 AND E-4 IN THEIR INITIAL TERM OF ENLISTMENT.

B. SIGNIFICANT FINDINGS AND CONCLUSIONS

1. THE FACTOR "MY WORK IS INTERESTING", ONE OF THE 17 ORIGINAL INDEPENDENT FACTORS IN THE AOSP MEASURED ON A FIVE POINT SCALE RANGING FROM "NONE OF THE TIME" TO "ALL OF THE TIME", EMERGED AS THE BEST PREDICTOR OF BOTH REENLISTMENT INTENT AND JOB SATISFACTION. THIS FINDING WAS NOTED FOR E-3'S AND E-4'S SEPARATELY, MALES AND FEMALES, NON-HIGH SCHOOL GRADUATES, HIGH SCHOOL GRADUATES, WHITES AND BLACKS, AND SINGLE AND MARRIED PERSONNEL. IT IS NOTED THAT THIS FACTOR (INTRINSIC TO ONE'S JOB) APPEARED TO EXERT MUCH MORE INFLUENCE ON REENLISTMENT INTENT AS WELL AS JOB SATISFACTION THAN FACTORS PERTAINING TO ONE'S CAREER, PARTICULARLY MONETARY-RELATED FACTORS COMPRISING MILITARY PAY, ALLOWANCES, AND BENEFITS. IN VIEW OF THE NEED OF THE ARMY TO REDUCE PERSONNEL-RELATED COSTS WHILE INCREASING THE RETENTION RATE OF QUALIFIED PERSONNEL, ESPECIALLY UNDER THE ALL VOLUNTEER

104

FORCE, MAKING JOBS MORE ATTRACTIVE COULD BE EXTREMELY DESIRABLE. IT WAS ALSO DETERMINED THAT THE EXPRESSED REENLISTMENT INTENT OF FIRST-TERM PERSONNEL WAS HIGHLY CORRELATED WITH ACTUAL REENLISTMENT DECISION. THIS WAS ESPECIALLY TRUE AS THEY APPROACHED THE DECISION POINT REGARDING REENLISTMENT. SIMILAR STUDIES CONDUCTED BY THE U.S. NAVY AND THE U.S. AIR FORCE ON FIRST-TERM PERSONNEL HAVE ALSO SHOWN VERY HIGH CORRELATIONS.

2. REGULAR MILITARY COMPENSATION (THE SUM OF BASIC PAY, QUARTERS AND SUBSISTENCE ALLOWANCES OR EQUIVALENT, AND FEDERAL INCOME TAX ADVANTAGE COMPARED TO SALARY/WAGES MADE IN CIVILIAN LIFE), NOT ONE OF THE ORIGINAL FACTORS USED IN THE AOSP, WAS GENERALLY A CONSISTENT PREDICTOR OF REENLISTMENT INTENT BUT TO A LESSER EXTENT THAN WORK INTEREST. THIS WAS TRUE REGARDLESS OF THE SOLDIER'S SEX OR RACE. THIS ALSO APPLIED TO E-4'S, SINGLE PERSONNEL, AND HIGH SCHOOL DEGREE GRADUATES, BUT NOT THEIR COMPLEMENTS.

3. WORK IMPORTANCE, WORK CHALLENGE, AND WORKING ASSOCIATION WITH ONE'S SUPERVISORS WERE RELATIVELY CONSISTENT PREDICTORS OF JOB SATISFACTION IN TERMS OF GRADE, SEX, EDUCATIONAL LEVEL, RACE, AND MARITAL STATUS.

4. SOLDIERS WHO FELT THEY WERE GIVEN ACCURATE INFORMATION BY THEIR ARMY RECRUITER HAD A SIGNIFICANTLY HIGHER INTENTION TO REENLIST AND HAD SIGNIFICANTLY GREATER JOB SATISFACTION THAN THOSE WHO DIDN'T. THE BELIEF THAT ARMY RECRUITERS TOLD THE TRUTH ABOUT ARMY LIFE DOES NOT NECESSARILY IMPLY THAT THEY EITHER TRULY REPRESENTED OR MISREPRESENTED THE FACTS ABOUT ARMY LIFE.

What this indicated was the extent to which the expectations of the individual correspond to the information imparted to him/her by the Army recruiter. Those individuals more likely to accept Army life for what it really is, regardless of the information by the recruiter, are in turn much more likely to reenlist and to be satisfied with their job.

III. The February 1977 Army-wide Survey

A. Introduction.

Just as for the April 1977 Pilot Test, because of time and manpower constraints it was decided to utilize in part the empirically developed job satisfaction factors from the AFHRL for a survey to constitute Phase II of the overall project. Other factors were added based on the previously described August 1976 Army-wide Survey and its analysis. Work conducted by the US Army Research Institute for the Behavioral and Social Sciences was carefully considered for application to the project. A report by N.W. Ayer, Inc. on the attitudes and motivations of first-termers toward reenlistment and a study done by the Office of the Deputy Chief of Staff for Personnel, Department of the Army, on the attitudes of soldiers leaving the Army were also researched. These efforts culminated in the development of an 80 item questionnaire administered Army-wide to a random sample of 3708 soldiers in February 1977. Forty-two of these items pertained directly to an evaluation of satisfaction on a seven point scale ranging from "Extremely Dissatisfied" to "Extremely Satisfied". The remaining questions provided

BACKGOUND INFORMATION AND ADDRESSED AREAS THOUGHT TO INFLUENCE JOB SATISFACTION OR REENLISTMENT INTENT BUT WHICH COULD NOT BE EFFECTIVELY MEASURED ON A SATISFACTION SCALE.

OF THE 3,708 CASES ON WHICH THE ANALYSIS WAS BASED, 1,532 COMPRISED THE FIRST TERM SAMPLE WHILE THE CAREER FORCE SAMPLE CONTAINED 2,176 INDIVIDUALS. ALL FIRST-TERM SOLDIERS WERE IN PAYGRADE E-5 OR BELOW AND HAD LESS THAN FOUR YEARS OF ACTIVE FEDERAL MILITARY SERVICE. ALL THE MEMBERS OF THE CAREER FORCE WERE SERVING A SECOND OR SUBSEQUENT ENLISTMENT; WERE IN PAYGRADE E-3 AND ABOVE; AND HAD AT LEAST THREE YEARS OF ACTIVE FEDERAL MILITARY SERVICE.

B. SIGNIFICANT FINDINGS AND CONCLUSIONS

(1) ASPECTS OF ARMY LIFE VIEWED AS THE MOST AND LEAST SATISFYING:

IN GENERAL, SOLDIERS INDICATED GREATEST SATISFACTION WITH FACTORS INTRINSIC TO THEIR WORK AND THE GREATEST DISSATISFACTION WITH EXTRINSIC OR SITUATIONAL FACTORS, AS INDICATED IN TABLES 2 AND 3. FOR EXAMPLE, FIRST-TERMERS WERE MOST SATISFIED WITH THE SECURITY PROVIDED BY THEIR JOBS WHILE CAREERISTS WERE MOST SATISFIED WITH THE OPPORTUNITY TO HELP OTHERS BY DOING THEIR JOB. ON THE OTHER HAND, BOTH GROUPS WERE LEAST SATISFIED WITH THE WAY THE ARMY MAKES USE OF ITS ENLISTED PERSONNEL. EXAMINATION OF THE RESPONSES FROM FIRST-TERM SUBGROUPS (E.G., MENN,WOMEN, HIGH SCHOOL DEGREE GRADUATES, NON-HIGH SCHOOL DEGREE GRADUATES) ALSO REVEALED SATISFACTION WAS LOWEST WITH REGARD TO PERSONNEL UTILIZATION. THIS WIDESPREAD SENSE OF MALUTILIZATION

WOULD ARGUE STRONGLY FOR ADDITIONAL SENSITIVITY BY THE ARMY TOWARD EFFECTIVE ASSIGNMENT AND USE OF ENLISTED PERSONNEL. INCREASED EFFORTS TO PROVIDE MEANINGFUL WORK, ENSURE THAT TRAINING IS A REFLECTION OF JOB REQUIREMENTS, AND IMPROVE THE ACTUAL MATCH BETWEEN PRIMARY MOS AND WORK PERFORMED WOULD BE MOST BENEFICIAL.

(2) PREDICTORS OF JOB SATISFACTION, REENLISTMENT INTENT, AND UNIT MORALE

(A) JOB SATISFACTION

AS INDICATED IN TABLE 4, SATISFACTION WITH WORK PERFORMED IN TERMS OF INTEREST, IMPORTANCE, CHALLENGE, VARIETY, AND THE USE OF TRAINING AND ABILITIES WAS THE PRIMARY PREDICTOR OF JOB SATISFACTION FOR BOTH FIRST-TERMERS AND CAREERISTS. SATISFACTION WITH THEIR SUPERVISOR'S LEADERSHIP, TECHNICAL AND ADMINISTRATIVE SKILLS WAS ALSO IMPORTANT TO BOTH GROUPS. FOR FIRST-TERMERS, CHANGES IN THE WORK PERFORMED HAS THE GREATEST POTENTIAL FOR IMPROVING THE ATTITUDES OF FIRST-TERM SOLDIERS. THIS CONCLUSION, HOWEVER, IS CONTINGENT ON PROVIDING WORK RELATED TO ONE'S PRIMARY MOS AND RELEVANT TO TRAINING RECEIVED.

COMPARED TO FIRST-TERMERS, THE OVERALL JOB SATISFACTION OF CAREERISTS WAS MORE CLOSELY ASSOCIATED WITH SATISFACTION TOWARD THEIR WORK SCHEDULES (RELATING TO THE LENGTH OF ONE'S WORK HOURS), OPPORTUNITIES FOR WORKING AND ASSOCIATING WITH PEOPLE THEY LIKE, AND HAVING RESPONSIBILITY FOR SEEING A JOB THROUGH TO COMPLETION.

(B) REENLISTMENT INTENT

AMONG THE "BEST" PREDICTORS OF REENLISTMENT INTENT

FOR FIRST-TERMERS AND CAREERISTS, AS OBSERVED IN TABLE 5, ONLY RELATIVE SATISFACTION WITH PAY AND ALLOWANCES EMERGED FOR BOTH GROUPS. FIRST-TERM SOLDIER'S ATTITUDES TOWARD THEIR WORK, THE ARMY'S USE OF ENLISTED PERSONNEL (A SIGNIFICANT PREDICTOR OF JOB SATISFACTION, AS PREVIOUSLY NOTED), AND RECRUITER ACCURACY WERE ALSO IMPORTANT CONTRIBUTORS TO REENLISTMENT PLANS. PERTAINING TO THE LATTER ASPECT, IT WOULD APPEAR THAT AN ACCURATE AND RELATIVELY COMPLETE PORTRAYAL OF ARMY LIFE AND WORK BY THE RECRUITER IS AN ESSENTIAL INGREDIENT FOR THE LONG RANGE RETAINABILITY OF FIRST-TERMERS (ALSO FOUND IN THE AUGUST 1976 ARMY-WIDE SURVEY AS PREVIOUSLY DISCUSSED).

ALONG WITH SATISFACTION TOWARD PAY AND ALLOWANCES, THE FACTORS IDENTIFIED AS THE "BEST" PREDICTORS OF REENLISTMENT INTENT FOR CAREER SOLDIERS WERE: ARMY POLICIES AND PROCEDURES (E.G., PROMOTION, EVALUATION, REENLISTMENT, DISCIPLINE); FAMILY RECOGNITION AND PRIDE IN THE SOLDIER'S WORK; AND DUTY LOCATION. AMONG THESE FACTORS, DISSATISFACTION WAS EXPRESSED ONLY WITH POLICIES AND PROCEDURES. THIS SITUATION DID NOT APPEAR TO BE A PROBLEM OF COMMUNICATION SINCE CAREERISTS WERE BASICALLY SATISFIED WITH THE AVAILABILITY OF INFORMATION CONCERNING FACETS OF ARMY LIFE.

(c) UNIT MORALE

THE "BEST" PREDICTORS OF UNIT MORALE FOR BOTH FIRST-TERM AND CAREER SOLDIERS, AS SHOWN IN TABLE 6, INCLUDED THE LEVEL OF SATISFACTION WITH: PRIDE THAT CO-WORKERS HAVE IN

THE UNIT AND ARMY; UNIT POLICIES AND PROCEDURES (E.G., PROMOTION, LEAVE, TIME-OFF, EVALUATION); AND TRAINING GIVEN AT UNIT LEVEL. SATISFACTION WITH THE QUALITY AND AVAILABILITY OF BOTH ON AND OFF-POST EATING FACILITIES, AND THE ARMY'S EMPHASIS ON EQUALITY OF THE SEXES WERE ALSO IMPORTANT TO PREDICTING THE OPINION OF FIRST-TERMERS OF UNIT MORALE. FOR CAREERISTS, RELATIVE SATISFACTION WITH THEIR SUPERVISORS' SKILLS AND THE AVAILABILITY OF NECESSARY INFORMATION CONCERNING UNIT POLICIES AND PROCEDURES ALSO CONTRIBUTED TO PREDICTING ATTITUDES TOWARD UNIT MORALE.

(3) ENLISTMENT AND SEPARATION REASONS

(A) ENLISTMENT

EXAMINATION OF THE INITIAL PLANS OF FIRST-TERM SOLDIERS TOWARD AN ARMY CAREER AT THE TIME OF ENLISTMENT IN CONJUNCTION WITH THEIR REASONS FOR ENLISTMENT SUGGESTS THESE THREE BASIC CATEGORIES: RECRUITS PLANNING TO SERVE ONLY ONE ENLISTMENT (33 PERCENT OF ALL FIRST-TERMERS); THOSE ENLISTING WITHOUT ANY CONCRETE IDEAS CONCERNING AN ARMY CAREER (ABOUT 40 PERCENT OF THE SAMPLE); AND THOSE WHO JOINED INTENDING TO MAKE THE ARMY A CAREER (COMPRISING ABOUT 20 PERCENT OF THE RECRUITS).

ENLISTMENT REASONS SELECTED BY SOLDIERS WERE GROUPED INTO FOUR CATEGORIES TO FACILITATE ANALYSIS: (1) ENLISTMENT OPTIONS/INCENTIVES; (2) NO PERSONAL COMMITMENT; (3) PATRIOTIC - ARMY INTRINSIC; AND (4) "OTHER". AS SHOWN IN TABLE 7, AMONG FIRST-TERMERS, ENLISTMENT OPTIONS/INCENTIVES, ACCOUNTING FOR 41.8 PERCENT OF ENLISTMENTS, WERE THE MOST COMMON REASONS SELECTED.

OF THOSE CHOOSING ENLISTMENT OPTIONS/INCENTIVES, NEARLY TWO-FIFTHS RESPONDED TO "GI EDUCATIONAL BENEFITS" OR "LEARNING A SKILL OR TRADE TO USE IN CIVILIAN LIFE" AS THEIR PRIMARY ENLISTMENT INDUCEMENT. ON THE OTHER HAND, THE LARGEST PERCENTAGE OF CAREER PERSONNEL (45.4 PERCENT) INDICATED THEY HAD ENTERED THE ARMY FOR PATRIOTIC/ARMY INTRINSIC REASONS; THIS PERCENTAGE WAS MARKEDLY HIGHER THAN THAT OF 26 PERCENT FOR FIRST-TERMERS. SUCH REASONS INCLUDED SERVICE TO THE COUNTRY AND THE CHANCE FOR ADVENTURE, TRAVEL AND NEW EXPERIENCES (THESE PARTICULAR REASONS ACCOUNTING FOR JUST OVER ONE-THIRD OF CAREERIST ENLISTMENTS).

THE ENLISTMENT REASONS OF FIRST-TERMERS WERE ALSO EXAMINED BASED ON THEIR INITIAL PLANS TOWARD AN ARMY CAREER. AS DISPLAYED IN TABLE 8, OF THOSE FIRST-TERM PERSONNEL PLANNING TO SERVE ONLY ONE TERM, THE MAJORITY PICKED REASONS CATEGORIZED AS ENLISTMENT OPTIONS/INCENTIVES AS THEIR PRIME MOTIVATORS FOR JOINING THE ARMY. WITHIN THESE REASONS, "GI EDUCATIONAL BENEFITS" AND "LEARNING A SKILL OR TRADE TO USE IN CIVILIAN LIFE" ACCOUNTED FOR APPROXIMATELY 50 PERCENT OF THIS GROUP'S ENLISTMENT. FIRST-TERMERS WHO HAD NO REAL PLANS CONCERNING AN ARMY CAREER AT THE TIME OF ENLISTMENT TENDED TO JOIN FOR REASONS CATEGORIZED AS "NO PERSONAL COMMITMENT" (E.G., TAKING TIME TO GROW-UP, GETTING AWAY FROM HOME TOWN, AND NEED FOR A JOB). RECRUITS WHO INITIALLY PLANNED TO MAKE THE ARMY A CAREER WERE MOST LIKELY TO CITE FACTORS ASSOCIATED WITH ENLISTMENT OPTIONS/INCENTIVES AS HAVING

CONTRIBUTED MOST TO THEIR JOINING. HOWEVER, THEY WERE ABOUT TWICE AS LIKELY AS EITHER OF THE OTHER TWO GROUPS TO HAVE ENLISTED DUE TO PATRIOTIC OR ARMY INTRINSIC REASONS.

THESE FINDINGS SUGGEST THERE IS A NEED FOR ADDITIONAL RECRUITING EMPHASIS ON HOW THE ARMY CAN CHALLENGE INDIVIDUALS (IN TERMS OF TRAINING, SERVICE TO THE NATION, DISCIPLINE, ADVENTURE, AND TRAVEL) SINCE THESE COMPONENTS HAVE THE POTENTIAL FOR ATTRACTING QUALITY RECRUITS WHO ARE FAR MORE LIKELY TO MAKE THE ARMY A CAREER. ON THE OTHER HAND, THE WIDESPREAD USE OF ENLISTMENT OPTIONS AND INCENTIVES BEYOND TRAINING AND EDUCATION (E.G., UNIT-OF-CHOICE, ARMY AREA/STATION-OF-CHOICE, CASH BONUS) COULD BE CURTAILED OR ELIMINATED WITH THE POTENTIAL FOR CONSIDERABLE DOLLAR SAVINGS AS WELL AS INCREASED ASSIGNMENT FLEXIBILITY.

(B) SEPARATION

REASONS FOR SEPARATION WERE CLUSTERED INTO FIVE CATEGORIES: (1) ARMY POLICIES/PROCEDURES/LIFE; (2) ONE-TERM OR SHORT-TERM MOTIVATIONS; (3) JOB RELATED; (4) PERSONAL MOTIVATION; AND (5) "OTHER". AS SHOWN IN TABLE 9, ABOUT TWO-FIFTHS OF THE FIRST-TERMERS AND CAREERISTS WHO DEFINITELY PLANNED TO SEPARATE TENDED TO SELECT FACTORS ASSOCIATED WITH ARMY POLICIES/PROCEDURES/LIFE AS HAVING MOST INFLUENCED THEIR DECISION TO LEAVE THE ARMY. THEY CITED THE AMOUNT OF BUSY WORK, HARASSMENT, AND EXTRA DUTIES; AND EXCESSIVE CONCERN FOR HAIRCUTS, APPEARANCE, AND DISCIPLINE AS THE MOST IMPORTANT REASONS FOR THEIR INTENDED SEPARATION. IN ADDITION, FIRST-TERM SOLDIERS ALSO IDENTIFIED

112

LOW PAY AND ALLOWANCES AS AN IMPORTANT CAUSE FOR SEPARATION WHILE CAREERISTS NOTED DISDAIN FOR THEIR CURRENT MOS AND BEING UNABLE TO GET ONE THEY WANTED AMONG FACTORS MOST INFLUENCING THEIR DECISIONS TO LEAVE THE ARMY.

THE PROPENSITY TO REENLIST AMONG FIRST-TERM SOLDIERS WHO ENTERED THE ARMY INTENDING TO SERVE ONLY ONE TERM APPEAR TO BE UNAFFECTED BY THEIR ARMY EXPERIENCES. AS INDICATED IN TABLE 10, THEY TENDED TO JOIN TO PURSUE SPECIFIC GOALS (E.G., GI EDUCATIONAL BENEFITS), AND HAVING ATTAINED THESE OBJECTIVES PREFER TO SEPARATE. THOSE INDIVIDUALS WHO JOINED WITHOUT ANY CLEAR-CUT PLANS TOWARD AN ARMY CAREER DECIDED TO SEPARATE BECAUSE OF EXCESSIVE CONCERN FOR HAIRCUTS, APPEARANCE AND DISCIPLINE AS WELL AS THE AMOUNT OF BUSY WORK, HARASSMENT AND EXTRA DUTIES. PERCEPTIONS OF HAVING VERY LITTLE "REAL WORK" TO DO WERE ALSO RESPONSIBLE FOR INCLINATIONS TOWARD SEPARATION FOR THIS GROUP. AMONG FIRST-TERMERS WHO INITIALLY DESIRED AN ARMY CAREER, LOW PAY AND ALLOWANCES WERE SELECTED AS CONTRIBUTING MOST TO THE DECISION OF SOLDIERS IN THIS GROUP TO SEPARATE. BUSY WORK, HARASSMENT AND EXTRA DUTY TOGETHER WITH THE ABSENCE OF "REAL WORK" WERE ALSO FREQUENTLY CITED REASONS.

ONLY TWO JOB OR WORK RELATED FACTORS WHICH CONTRIBUTE SIGNIFICANTL TO A SEPARATION DECISION (AMOUNT OF BUSY WORK, HARASSMENT, AND EXTRA DUTIES; AND TOO LITTLE "REAL WORK" TO DO) APPEAR TO BE ADDRESSABLE BY THE ARMY. ALTHOUGH OBVIOUS, PROVIDING SOLDIERS WITH INTERESTING WORK WHICH CHALLENGES THEIR TALENTS AND training promises to create an environment more conducive to reenlistment (also indicated in the August 1976 survey). In particular, an increase in meaningful work will raise overall job satisfaction, heighten reenlistment intent, and ultimately increase reenlistment.

IV. The April 1977 Pilot Test

A. Introduction.

To provide the best possible coverage of those factors which could be used to assess the inter-relationships between reenlistment decision, unit morale, and job/career satisfaction, a pilot test questionnaire was developed over a period of three months. This questionnaire represented the transition from the Herzberg-based appraoch utilized in the job satisfaction portion of the AOSP to an eclectic approach combining the work of the Army Research Institute (ARI), the Air Force Human Resources Laboratory (AFHRL), and MILPERCEN. Although it had been intended to develop and test a job and career satisfaction model wholly within MILPERCEN, because of time and manpower constraints it was decided to capitalize on the extensive literature review and long-range research conducted by the AFHRL on job/career satisfaction.

Consequently, the items used in the pilot test questionnaire were derived in large measure on an Occupational Attitude Inventory (OAI) developed by the AFHRL. In the initial development of the OAI, 36 potential satisfaction dimensions or

HYPOTHESIZED FACTORS WERE IDENTIFIED BY AIR FORCE BEHAVIORAL SCIENTISTS FAMILIAR WITH THE MILITARY WORK ENVIRONMENT. ITEMS WERE WRITTEN FOR EACH DIMENSION, RESULTING IN A FINAL POOL OF 348 ITEMS (APPROXIMATELY 10 ITEMS PER DIMENSION) WHICH WERE VALIDATED THROUGH ANALYSIS OF THE RESPONSES BY A RANDOM SAMPLE OF ABOUT 3,000 FIRST-TERM AIRMEN.

FOR USE IN AN ARMY ENVIRONMENT, AS INDICATED IN TABLE 11, 32 OF THE HYPOTHESIZED FACTORS IN THE OAI WERE MODIFIED WHILE FOUR NEW FACTORS WERE ADDED (ENTITLED, "FAMILY", "INDIVIDUAL", "DISCRIMINATION", AND "ARMY UNIQUE"). IT WAS BELIEVED THAT THESE ADDITIONAL FACTORS REPRESENT IMPORTANT INFLUENCES ON A PERSON'S MOTIVATION AND BEHAVIOR. OF A TOTAL POOL OF 324 ITEMS SELECTED INITIALLY IN PILOT TEST, 225 WERE RETAINED. REDUCTION OF THE NUMBER OF ITEMS WAS BASED ON THE FOLLOWING CRITERIA:

(1) REDUNDANCY

(2) REDUCING THE EXCESSIVELY LARGE NUMBER OF ITEMS IN THE FACTORS ENTITLED "INDIVIDUAL", "HUMAN SUPERVISION" AND "FAMILY".

THE PILOT TEST QUESTIONNAIRE WAS ADMINISTERED TO APPROXIMATELY 1,600 PERSONNEL IN APRIL 1977 AT SIX CONUS INSTALLATIONS. IN ADDITION, ABOUT 600 SOLDIERS WERE INTERVIEWED, PRIMARILY TO PROVIDE INSIGHTS INTO THE CONTENT VALIDITY OF THE QUESTIONNAIRE AND CLARITY OF INSTRUCTIONS.

THE FINAL INSTRUMENT WAS REDUCED FROM 225 TO 124 ITEMS THROUGH USE OF FACTOR ANALYSIS, STEPWISE MULTIPLE REGRESSION ANALYSIS, AND A SUBJECTIVE REVIEW. THE SUBJECTIVE REVIEW WAS USED TO ELIMINATE DUPLICATION WITHIN EACH OF THE HYPOTHESIZED FACTORS AND TO ELIMINATE ITEMS JUDGED TO BE OF LITTLE PRACTICAL VALUE IN TERMS OF JOB/ARMY CAREER SATISFACTION, UNIT MORALE, AND RETENTION SUCH AS "YOUR OPINION OF THE ARMY COMPARED TO THE AIR FORCE". THE 124 ITEMS THEN CONSTITUTED ALL THE ITEMS COMPRISING SECTION B OF THE COMPREHENSIVE ARMY-WIDE JOB AND CAREER SATISFACTION SURVEY ADMINISTERED IN NOVEMBER 1977.

116

TABLE 1

FACTORS USED IN THE AOSP

I. INDEPENDENT FACTORS

"HYGIENE"	"MOTIVATOR"
Work Interest	Work Conditions (Facilities, Equipment Tools)
Recognition Received For Work Done	Quality Of Technical Supervision Received
Work Variety	Job Prestige (How Job Ranks With Other Soldier's Jobs)
Opportunity To See Work Results	Conflict Of Job With Family Responsibilities
Work Importance	Working Association With Co-Workers
Amount Of Responsibility	Working Association With Supervisors
Opportunity To Increase Job Skills And Knowledge	Army Pay (Base Pay, Allowances, Special Pay)
Opportunity For Promotion, Increase In Job Status	Army Benefits (PX, Commissary, Medical)
Work Challenges Training, Skills, Knowledge	

II. CRITERION MEASURES

Satisfaction With Present Job

Reenlistment Plans

TABLE 2

FACTORS WITH WHICH FIRST-TERM AND CAREER FORCE PERSONNEL ARE MOST SATISFIED

FACTOR	1ST TERM PERSONNEL (RANK)	CAREER PERSONNEL (RANK)
CHANCE TO HELP OTHERS BY DOING JOB	1	3
CHANCE TO HAVE RESPONSIBILITY FOR SEEING A JOB THROUGH TO COMPLETION	2	-
OPPORTUNITY TO WORK AND ASSOCIATE WITH PEOPLE YOU LIKE	3	2
JOB SECURITY	4	1
PRIDE YOUR FAMILY HAS AND RECOGNITION YOUR FAMILY GIVES TO YOUR JOB	5	4
AVAILABILITY OF ON-POST FACILITIES	-	5

118

TABLE 3

FACTORS WITH WHICH FIRST-TERM AND CAREER FORCE PERSONNEL ARE LEAST SATISFIED

FACTOR	1ST TERM PERSONNEL (RANK)	CAREER PERSONNEL (RANK)
THE WAY THE ARMY UTILIZES ENLISTED PERSONNEL	1	1
THE WAY THE ARMY MAKES USE OF EQUIPMENT, MATERIAL, SUPPLIES	2	3
QUALITY AND AVAILABILITY OF HOUSING (ON AND OFF-POST)	3	2
STANDARD OF LIVING YOU NOW HAVE	4	4
"RED TAPE" ASSOCIATED WITH YOUR JOB	5	-
PRIDE YOUR CO-WORKERS HAVE IN YOUR UNIT AND THE ARMY	-	5

TABLE 4

"BEST" FIVE PREDICTORS OF JOB SATISFACTION FOR FIRST-TERM AND CAREER PERSONNEL

FACTOR	FIRST-TERM PERSONNEL (RANK)	CAREER PERSONNEL (RANK)
PRESENT DUTIES (CHALLENGE, INTEREST, IMPORTANCE)	1	1
CHANCE TO ACQUIRE TRAINING, EXPERIENCE, SKILLS, AND KNOWLEDGE WHICH CAN BE USED IN A CIVILIAN JOB	2	-
YOUR SUPERVISOR'S LEADERSHIP, TECHNICAL AND ADMINISTRATIVE SKILLS	3	3
CHANCE TO HELP OTHERS BY DOING YOUR JOB	4	-
AMOUNT OF WORK YOU HAVE TO DO	5	-
WORK SCHEDULE (TOTAL HOURS, SHIFTS, PACE OF WORK)	-	2
OPPORTUNITY TO WORK AND ASSOCIATE WITH PEOPLE YOU LIKE	-	4
CHANCE TO HAVE RESPONSIBILITY FOR SEEING A JOB THROUGH TO COMPLETION	-	5

120

TABLE 5

"BEST" FIVE PREDICTORS OF REENLISTMENT INTENT FOR FIRST-TERM AND CAREER PERSONNEL

FACTOR	FIRST-TERM PERSONNEL (RANK)	CAREER PERSONNEL (RANK)
PRESENT DUTIES (CHALLENGES, INTEREST, IMPORTANCE)	1	-
ARMY PAY AND ALLOWANCES	2	2
IN GENERAL, THE THINGS THE RECRUITER TOLD ME ABOUT THE ARMY WERE TRUE	3	-
THE WAY THE ARMY UTILIZES ENLISTED PERSONNEL	4	-
DOING WORK WHICH BOTHERS YOUR CONSCIENCE	5	-
ARMY POLICIES AND PROCEDURES	-	1
PRIDE YOUR FAMILY HAS AND RECOGNITION YOUR FAMILY GIVES TO YOUR JOB	-	3
DUTY LOCATION	-	4
YEARS OF ACTIVE FEDERAL MILITARY SERVICE	-	5

121

TABLE 6

"BEST" FIVE PREDICTORS OF UNIT MORALE FOR FIRST-TERM AND CAREER PERSONNEL

FACTOR	FIRST-TERM PERSONNEL (RANK)	CAREER PERSONNEL (RANK)
Pride Your Co-Workers Have In Your Unit And The Army	1	1
Unit Policies And Procedures (Promotion, Evaluation, Leave, Training)	2	5
Quality And Availability Of Eating Facilities	3	-
Training Given In Your Unit	4	4
Amount Of Emphasis On Equality Of the Sexes	5	-
Your Supervisor's Leadership, Technical And Administrative Skills	-	2
Availability Of Necessary Information About Unit Policies and Practices (Promotion, Evaluation, Leave, Training)	-	3

122

TABLE 7

PERCENT OF ENLISTMENT REASONS BY CATEGORY FOR FIRST-TERM AND CAREER PERSONNEL

ENLISTMENT CATEGORY/REASONS	FIRST-TERM PERSONNEL %	CAREER PERSONNEL %
1. ENLISTMENT OPTIONS - INCENTIVES	41.8	23.1
A. To Become Eligible For GI Educational Benefits	19.3	5.4
B. To Learn A Skill/Trade To Use In Civilian Life	17.9	10.9
C. The Training Of Choice Option That I Wanted Was Available	2.1	2.4
D. The Enlistment Cash Bonus Was Available To Me	1.1	1.7
E. The Army Area/Station Of Choice Option That I Wanted Was Available	1.1	1.8
F. The Unit Of Choice Option That I Wanted Was Still Available	0.3	0.9
2. NO PERSONAL COMMITMENT	27.2	23.6
A. To Take Time Out To Find Myself, Grow-up, Mature	14.2	10.6
B. I Couldn't Get A Job (Or A Job I Wanted) Anywhere Else	5.9	6.9
C. To Get Away From My Home Town	5.8	5.4
D. I Had Friends Joining The Army Or Already In The Army	1.3	0.7

123

3. PATRIOTISM - ARMY INTRINSIC	26.0	45.4
A. The Chance For Adventure, Travel, And New Experiences	14.4	15.1
B. To Serve My Country	6.9	17.4
C. I Wanted To Be A Soldier	2.6	10.1
D. My Family Had A History Of Army Or Other Military Service	2.1	2.8
4. OTHER REASON - NOT LISTED	5.0	7.9

TABLE 8

PERCENT OF ENLISTMENT REASONS BY CATEGORY FOR FIRST-TERM SOLDIERS BY INITIAL PLANS CONCERNING AN ARMY CAREER

	ONE TERM	ARMY CAREER	NO PLANS
ENLISTMENT CATEGORY/REASONS	%	%	%
1. ENLISTMENT OPTIONS - INCENTIVES	54.1	42.8	37.5
A. To Become Eligible For GI Educational Benefits	35.7	19.2	19.8
B. To Learn A Skill/Trade To Use In Civilian Life	14.6	14.2	14.6
C. The Training Of Choice Option That I Wanted Was Available	1.0	3.9	1.5
D. The Enlistment Cash Bonus Was Available To Me	1.5	1.6	--.-
E. The Army Area/Station Of Choice Option That I Wanted Was Available	1.3	2.5	1.6
F. The Unit Of Choice Option That I Wanted Was Still Available	--.-	1.4	--.-
2. NO PERSONAL COMMITMENT	23.5	20.4	38.7
A. To Take Time Out To Find Myself, Grow-Up, Mature	11.2	14.1	18.1
B. I Couldn't Get A Job (Or A Job I Wanted) Anywhere Else	5.1	4.0	7.7
C. To Get Away From My Home Town	5.6	2.3	10.7
D. I Had Friends Joining The Army Or Already In The Army	1.6	--.-	2.2

3. PATRIOTISM - ARMY INTRINSIC	13.4	29.2	16.9
A. The Chance For Adventure, Travel, And New Experiences	9.8	7.6	13.0
B. To Serve My Country	2.0	7.8	2.2
C. I Wanted To Be A Soldier	--.-	7.1	0.6
D. My Family Had A History Of Army Or Other Military Service	1.6	6.7	1.1
4. OTHER REASON - NOT LISTED	9.0	7.5	7.0

TABLE 9

PERCENT OF SEPARATIONS BY CATEGORY/REASONS FOR FIRST-TERM AND CAREER SOLDIERS INDICATING THEY DEFINITELY PLAN TO SEPARATE

SEPARATION CATEGORY/REASONS	FIRST-TERM PERSONNEL %	CAREER PERSONNEL %
1. ARMY POLICIES/PROCEDURES/LIFE	37.9	45.9
A. I Think There Is Too Much Concern For Such Things As Haircuts, Appearance, And Discipline	9.8	8.3
B. The Pay And Allowances Are Too Low	9.8	6.3
C. The Amount Of Busy Work, Harassment And Extra Duties	9.6	11.6
D. I Don't Like My MOS And I Can't Arrange To Get One I Do Like	3.4	6.9
E. I Am Not Eligible To Reenlist	1.4	4.7
F. I Don't Think My Promotion Chances Are Too Good	1.4	3.1
G. I Couldn't Get The Reenlistment Option I Wanted	1.2	0.4
H. I Was Reclassified Into An MOS That I Have No Interest In And Don't Enjoy Working In	1.0	3.3
I. The Medical/Dental Care Is Inadequate	0.3	1.3
2. ONE-TERM/SHORT-TERM MOTIVATIONS	30.7	13.8
A. I Joined To Become Eligible For GI Educational Benefits	9.7	4.0
B. I Did Not Intend To Serve More Than One Enlistment	8.9	0.7
C. I Joined To Learn A Skill/Trade To Use In Civilian Life And I Have Done That	6.4	4.0

D. I Joined The Army To Have A Chance To Find Myself/Grow Up/Mature And I've Done That	4.7	4.9
E. I Joined The Army For Adventure/Travel/New Experiences And I've Accomplished These Things	1.0	0.2
3. JOB RELATED	19.3	20.2
A. I Think There Is Very Little "Real Work" To Do In The Army	9.5	5.9
B. I Spend Too Much Time Working Outside Of My Primary MOS	3.8	4.7
C. The Army Does Not Challenge Or Demand Enough Of Me	2.2	5.4
D. The Duty Hours Are Too Long And/Or Irregular	1.9	1.9
E. I Don't Like The People I Work For	1.9	2.3
4. PERSONAL MOTIVATIONS	7.6	11.7
A. My Wife/Husband Wants Me To Get Out	3.1	3.4
B. I Don't Like The People I Have To Associate With	2.2	2.6
C. My Living Conditions (Housing/Barracks) Are Poor	1.2	2.1
D. The Things I Can Gain From A Second Or Subsequent Enlistment (Job Training, Travel) Are Not Important Enough To Me	1.1	3.6
5. OTHER REASON-- NOT LISTED	4.6	8.4

TABLE 10

PERCENT OF SEPARATIONS BY CATEGORY/REASONS FOR FIRST-TERM SOLDIERS BY INITIAL PLANS CONCERNING AN ARMY CAREER WHO INDICATED THEY DEFINITELY PLAN TO SEPARATE

SEPARATION CATEGORY/REASONS	ONE TERM	ARMY CAREER	NO PLANS
	%	%	%
1. ARMY POLICIES/PROCEDURES	28.9	54.1	42.3
A. I Think There Is Too Much Concern For Such Things As Haircuts, Appearance, And Discipline	9.3	7.0	12.1
B. The Pay And Allowances Are Too Low	10.0	21.1	4.9
C. The Amount Of Busy Work, Harassment, And Extra Duties	5.8	17.3	11.8
D. I Don't Like My MOS And I Can't Arrange To Get One I Do Like	2.3	1.7	5.4
E. I Am Not Eligible To Reenlist	0.6	1.4	1.5
F. I Don't Think My Promotion Chances Are Too Good	0.4	1.7	2.2
G. I Couldn't Get The Reenlistment Option I Wanted	0.5	1.7	2.1
H. I Was Reclassified Into An MOS That I Have No Interest In And Don't Enjoy Working In	--.-	1.3	2.3
I. The Medical/Dental Care Is Inadequate	--.-	0.9	--.-
2. ONE-TERM MOTIVATIONS	45.2	6.0	21.3
A. I Joined To Become Eligible For GI Educational Benefits	15.2	0.7	6.7
B. I Did Not Intend To Serve More Than One Enlistment	16.2	1.3	2.7
C. I Joined To Learn A Skill/Trade To Use In Civilian Life And I Have Done That	6.7	2.2	6.3

D. I Joined The Army To Have A Chance To Find Myself/Grow Up/ Mature And I've Done That	5.4	0.5	5.5
E. I Joined The Army For Adventure/ Travel/New Experiences And I've Accomplished These Things	1.7	1.3	0.1
3. JOB RELATED	15.1	27.7	21.9
A. I Think There Is Very Little "Real Work" To Do In The Army	9.2	10.8	9.3
B. I Spend Too Much Time Working Outside Of My Primary MOS	2.5	6.8	4.1
C. The Army Does Not Challenge Or Demand Enough Of Me	0.9	5.1	3.1
D. The Duty Hours Are Too Long And/ Or Irregular	1.1	5.0	2.0
E. I Don't Like The People I Work For	1.4	--.-	3.4
4. PERSONAL MOTIVATIONS	7.2	4.6	9.5
A. My Wife/Husband Wants Me To Get Out	2.4	4.6	3.5
B. I Don't Like The People I Have To Associate With	1.6	--.-	3.5
C. My Living Conditions (Housing/ Barracks) Are Poor	1.6	--.-	1.5
D. The Things I Can Gain From A Second Or Subsequent Enlistment (Job Training, Travel) Are Not Important Enough To Me	1.6	--.-	1.0
5. OTHER REASON - NOT LISTED	2.6	7.6	5.0

TABLE 11

AIR FORCE		ARMY	
FACTOR DESCRIPTOR	NUMBER OF ITEMS	FACTOR DESCRIPTOR	NUMBER OF ITEMS
Achievement	7	Achievement	2
Activity	8	Activity	4
Air Force and Unit Policies and Practices	18	Army and Unit Policies and Practices	17
Assignment Locality	17	Assignment Locality	16
Authority	4	Authority	3
Co-workers	9	Co-workers	12
Creativity	10	Creativity	5
Importance	8	Importance	2
Interest	9	Interest	4
Knowledge of Results	7	Knowledge of Results	3
Personal Growth and Development	9	——	
Job Design	10	Job Design	3
Optional Social Contact Required Social Contact	7 10	Social Contact	11
Pay and Benefits	12	Pay and Benefits	8
Physical Work Environment	13	Physical Work Environment	9
Promotion Opportunity	8	Promotion Opportunity	4
Recognition	9	Recognition	4
Responsibility	10	Responsibility	4
Independence	9	——	
Value of Experience	8	Value of Military Experience	3

AIR FORCE		ARMY	
FACTOR DESCRIPTOR	NUMBER OF ITEMS	FACTOR DESCRIPTOR	NUMBER OF ITEMS
PHYSICAL SAFETY	6	PHYSICAL SAFETY	3
ECONOMIC SECURITY	4	ECONOMIC SECURITY	2
SERVICE TO OTHERS	8	SERVICE TO OTHERS	1
SOCIAL STATUS	11	SOCIAL STATUS	5
SUFFICIENCY OF TRAINING	12	SUFFICIENCY OF TRAINING	10
SUPERVISION RECEIVED - HUMAN RELATIONS	15	HUMAN SUPERVISION	16
SUPERVISION RECEIVED - TECHNICAL	9	TECHNICAL SUPERVISION	5
PERFORMANCE EVALUATION	8	PERFORMANCE EVALUATION	3
JOB CHANGE	7	JOB CHANGE	4
TOOLS, EQUIPMENT AND SUPPLIES	8	TOOLS, EQUIPMENT, AND SUPPLIES	7
UTILIZATION	8	UTILIZATION	3
VARIETY	9	VARIETY	4
WORK SCHEDULE	15	WORK SCHEDULE	6
SUPERVISORY DUTIES	18	SUPERVISORY DUTIES	10
UNCLASSIFIED	8	———	
———		INDIVIDUAL	14
———		ARMY UNIQUE	10
———		DISCRIMINATION	9
———		FAMILY	18
TOTAL	348	TOTAL	244

132

SECTION 2

OCCUPATIONAL-TASK ANALYSIS

PAPER PRESENTED AT THE 20TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION 1978

ORGANISATION: NAVAL MANPOWER UTILISATION UNIT,
HMS VERNON, PORTSMOUTH, ENGLAND

SUBJECT: EXECUTION OF LARGE OCCUPATIONAL ANALYSIS OF THE ROYAL NAVY'S OPERATIONS BRANCH

SPEAKER: MR C D BEEL

1. INTRODUCTION

SLIDES

Since 1971 RN Occupational Analysis has been carried out by the
Naval Manpower Utilisation Unit (NMUU). In 1973, through the generous
help of the US Navy, the use of the CODAP suite of computer programs was
obtained. Located at Portsmouth, Hampshire, the NMUU is an outport of 1
the Ministry of Defence Naval Manpower and Training Department.

It is staffed by a Commander in Charge, 5 Officers and 12 Chief Petty Officers with a small clerical staff.

2. THE OPERATIONS BRANCH

Until 1975 the various non technical enlisted men of the weapons, sensor, and communications operator branches of the RN were quite separate with their own structure and training organisation.

To increase efficiency and co-ordination in the modern warfare
environment these various independent branches were merged into sub-
branches of a new Operations Branch. This was to match the radical
changes made to the officer structure, including the introduction of the 2
Principal Warfare Officer trained to control the integrated fighting
systems of a ship. The School of Maritime Operations was set up as a
common faculty for Operations Branch and Principal Warfare Officer
Training.

It was decided to conduct an occupational analysis of the 11,000 men in the Branch during 1977 to see whether experience gained since its formation indicated any need for adjustment to training, duties, and structure.

There were several underlying reasons for the survey. Amongst the most important were:-

a. Concern that the new structure might lead to a loss of deep sub-specialist knowledge.
b. The need to establish how well the Branch was coping with man-
power shortages, shorter enlistment engagements and improved sea 3
shore ratios.
c. Whether further streamlining of training could be carried out.
d. Concern about retention of seamanship skills and the need for research into how this area of work was being apportioned between the sub-branches.

3. THE SURVEY OBJECTIVE

SLIDES

The survey objective was primarily for manpower structure and planning purposes with a spin off for Training Design. As in all recent NMUU surveys it aimed also to gather attitudes and opinions on many aspects of Service Conditions and job satisfaction. It was decided that data should be gathered not only from the job incumbent but also by a secondary questionnaire from his supervisors and managers.

To clarify beforehand what specific reports should be derived from 4
the data, and hence the questionnaire structure, the directive contained 5
very specific primary and secondary objectives.

The Survey occupied the entire resources of the NMUU and a considerable expenditure in computer processing over 18 months. A 73% sample (about 8500 men afloat and ashore) and 650 supervisors responded to their respective questionnaires.

Because of time constraints the remainder of my talk will be principally concerned with the main survey of surface Fleet ratings at sea and ashore.

4. QUESTIONNAIRE CONSTRUCTION 6

Information for the task inventory was gathered from every possible source, documentary and interview. The pilot fact finding survey sampled every sub branch and rate to cover as many different jobs as possible, by ship class. Over 500 people were interviewed using pre-planned data forms to obtain information at the job (rather than task) level under the broad headings:

Background Information

Billets

Qualifications

Ship Employment 7

Primary Work Area

Secondary Work Area

Work Area at Different Conditions

General Naval Duties

Seamanship Topics.

The information gathered was carefully collated and integrated with other sources of data.

Starting initially in specialist groups, then combining, a scalar diagram of all tasks of the Operations Branch was built up. These are 2 examples. From this was derived the basic task inventory to which were added specific questions needed to satisfy all aspects of the directive. 8

To meet the requirement to examine common operator and training 9
areas, the decision was taken to create one raw data base covering all sub branches based on one questionnaire structure.

SLIDES

The result was a formidable sized questionnaire - which created a dilemma. In practical terms we doubted whether any respondent could be asked to study every task in the inventory and all secondary questions without losing interest and producing dubious results. As you see by this example, we tried to keep him in the right frame of mind! But we did not want to constrain his answers into specific areas because of the commonality research aspect. Incidentally, as a matter of policy, the questionnaire is anonymous. 10

In the event a compromise was used on the task inventory, by subjectively dividing tasks into categories:

a. Pure Specialist

b. Common Ship Work 11

c. Areas of Likely Overlap.

This enabled us to limit the task of the respondent and hopefully to achieve the objective. Supporting information was gathered in similar categories.

5. COMPUTER FILE CONSTRUCTION

Two sections (Operational Duties and the task inventory) were incorporated in the questionnaire, both covering the whole man's apportionment of time but at different scalar levels. This arrangement could not be handled by CODAP in 1 computer file, so 2 CODAP files were envisaged and designed as part of questionnaire development. This carried with it some secondary benefits:

a. Operational Duties time section with service conditions/job satisfaction data only, reduced file length and computer times for this type of information. 12
b. Some CODAP internal size limits could be side stepped.
c. Operational Duties could be used as population identifiers in the main file for job descriptions from the task inventory.

6. EXECUTION

a. Public Relations

Based on our earlier experience and because the questionnaire had to be so large, and because our population extended over a large range of I.Q. and ability, an extensive publicity campaign was adopted.

(1) Several articles were published in the RN newspaper 'NAVY NEWS'.
(2) An authoritative instruction was issued. 13
(3) Letters were sent to all Commanding Officers.
(4) The foreword to the Questionnaire was signed by a Vice Admiral, Director General Naval Manpower and Training.

Additionally liaison visits were made to as many as possible ships and establishments in the UK by job analysts from the NMUU.

b. Distribution and return of Questionnaires — SLIDES

The despatch of questionnaires and their subsequent return gave to the Unit the aspect of a mail order business for the period June - September 1977. Some 11,000 were sent out. 14 (photograph)

In the event some 8500 questionnaires were returned. They were scrutinised to discard bad books. (This was a very low percentage, less than 1%).

Manual coding was limited to allocating 15 (photograph)

(1) Case Number
(2) Ship/Establishment Code
(3) 3 Digit Sub Branch/Rate

More refined ship/establishment coding providing various types of classification for grouping them was done by computer program, to reduce human error.

c. Data Capture

Data capture by optical mark reading would have been preferred but was not available on cost grounds.

Key to disk processing was used at the RNs Bureau West facility. 16 (photograph)

Data was transferred by tape to the computer and was programmed (as one combined operation) to cumulatively build a SEA and SHORE file in the following stages:-

(1) Coding Checks 17
(2) Refined Coding Additon
(3) Sorted into Sub Branch/Rate Order
[illegible] Merged in Sort Order into File.

Despite all the checks, some 'rogue' cases were not detected at this time and later caused problems of denigration of output.

d. Computer Processing

Final file statistics were:

SEA	5800	Each of 39 card images
SHORE	2200	record length
SUBMARINE	820	14 Card images.

Processing of NMUU CODAP is by batch mode at an Army Pay Computer and is run when time allows between primary pay processing. The size of the SEA file in particular created elapsed time problems which had not been fully anticipated. Some job runs went to 8 hours elapsed time. Many were stopped by the operator because of other requirements. To meet this a policy of splitting work had to be adopted to enable part jobs to run in smaller time gaps between pay runs - but this led to analysis problems at desk level.

SLIDES

A decision to produce a complete package report integrating all aspects for each work area or particular group, eg. job description, incumbent attitudes to training, supervisor opinion on training, incumbent job satisfaction, etc, meant that many separate computer jobs had to be run to satisfy primary analysis needs for one report. Often the report was held up for one aspect whose computer job was waiting in line behind many more. Allocating priorities became the name of the game. Keeping records of what printouts had already been obtained became difficult.

e. Analysis

Because of the policy of integrated reports and a very limited distribution of raw printout, the entire unit work force has been involved in the production of reports, for the best part of 1 year. We would like to make a more direct use of the printouts by giving them a wider distribution to our 'customers'. But people generally seem to be in some way deterred by computer prints, and find them difficult to understand.

Despite the difficulties mentioned, 50 very useful reports have gone to the authorities interested. Here are a few examples to illustrate the variety and scope. 18

The NMUU can only point out significant data results and possible conclusions, but has no executive authority to decide what needs to be done. On the whole this is thought to be the best arrangement for a management information service like ours. But benefits will take some time to appear - NMUU reports provide only one contribution to management decision.

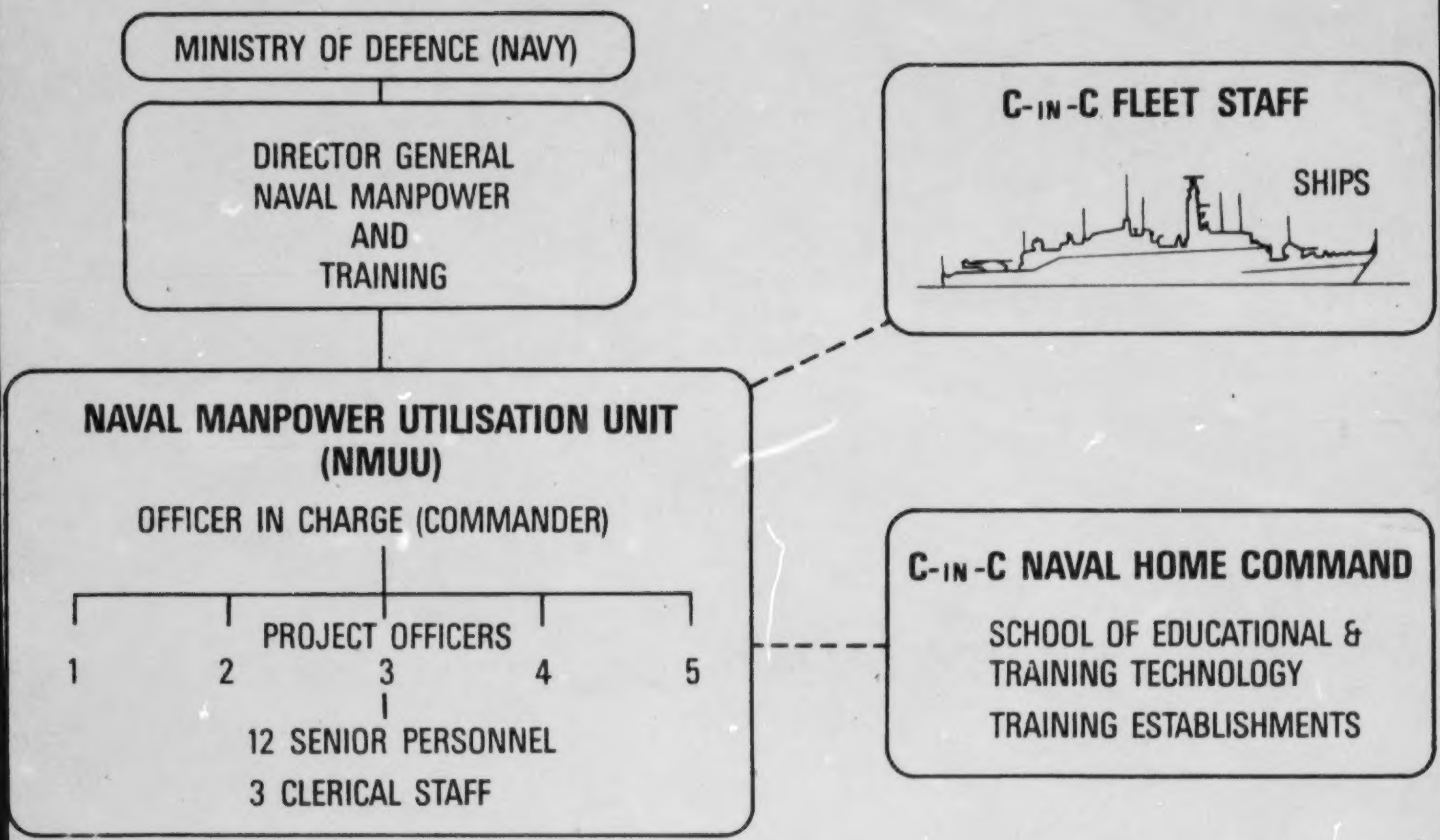
7. CONCLUSION

The Operations Branch Survey was a success in its planning, execution and results. The CODAP program package coped easily with the large files and did all that was expected of it. Nevertheless a few lessons were learned:-

a. Big is not necessarily beautiful. The sheer size of the job created many problems for a small Unit.
b. Mixing aims - Manpower, Training, Job Sat seems attractive in terms of a single visitation to the Fleet. But the value of results is downgraded by incompatibility of aims.
c. NMUU Staff suffered from lack of variety of work during a long analysis period. Some members joined after the survey started and barely saw the end of it. One could say that our own job satisfaction suffered a little!

Thankyou gentlemen. Before attempting to answer any questions may I say how much the Royal Navy and my Unit appreciate the privilege of attending this annual meeting where so many experienced authorities in the field are convened.

SLIDE 1



MINISTRY OF DEFENCE (NAVY)

DIRECTOR GENERAL
NAVAL MANPOWER
AND
TRAINING

NAVAL MANPOWER UTILISATION UNIT (NMUU)

OFFICER IN CHARGE (COMMANDER)

PROJECT OFFICERS

1 2 3 4 5

12 SENIOR PERSONNEL

3 CLERICAL STAFF

C-IN-C FLEET STAFF

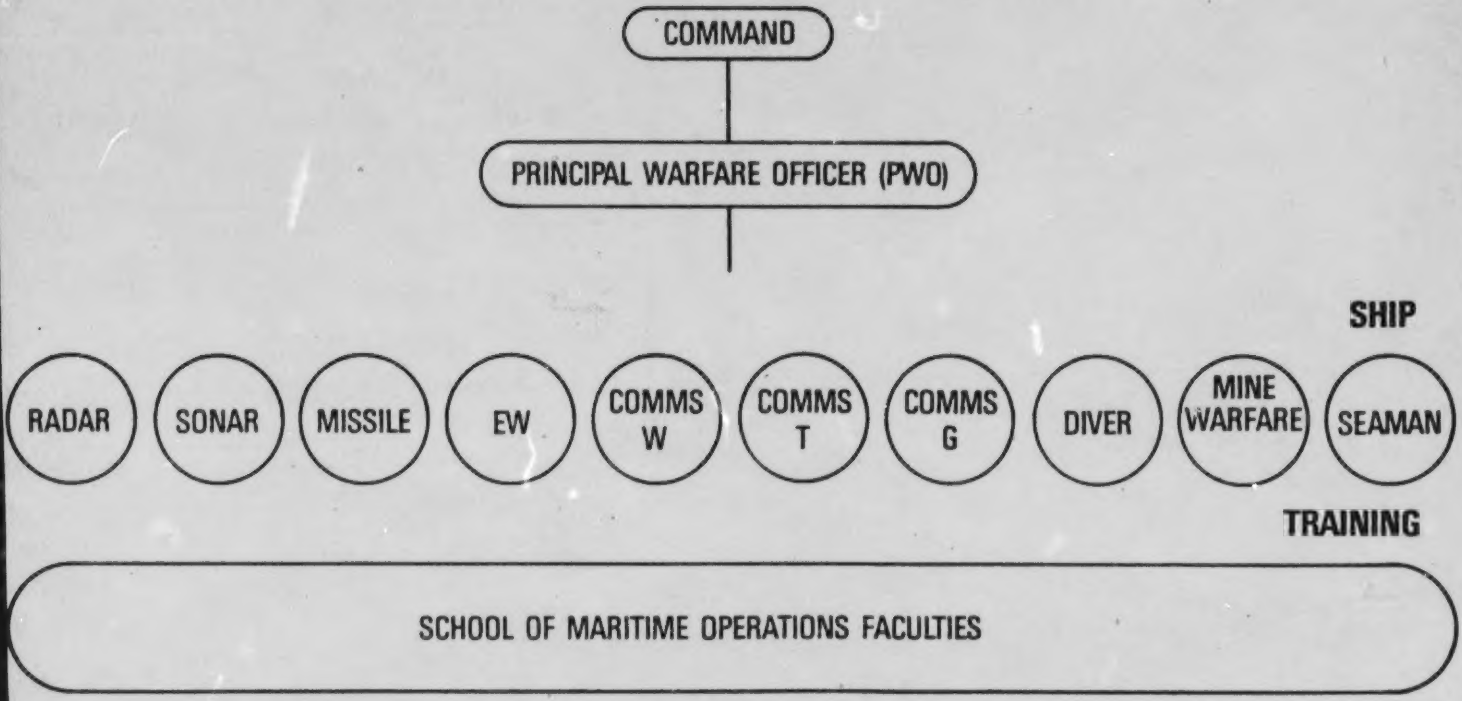
SHIPS

C-IN-C NAVAL HOME COMMAND

SCHOOL OF EDUCATIONAL & TRAINING TECHNOLOGY

TRAINING ESTABLISHMENTS

THE OPERATIONS BRANCH



COMMAND

PRINCIPAL WARFARE OFFICER (PWO)

SHIP

RADAR

SONAR

MISSILE

EW

COMMS W

COMMS T

COMMS G

DIVER

MINE WARFARE

SEAMAN

TRAINING

SCHOOL OF MARITIME OPERATIONS FACULTIES

OPERATIONS BRANCH SURVEY

1. LOSS OF DEEP SUB-SPECIALIST KNOWLEDGE
2. REDUCTION IN SHORE TRAINING/INCREASED SHIP TRAINING
3. MANPOWER SHORTAGES
4. IDENTIFICATION OF COMMON OPERATOR TASKS
5. SEAMANSHIP WORK

EXTRACT FROM OPERATIONS BRANCH DIRECTIVE TO NMUU

PURPOSE OF THE SURVEY

PRIMARY PURPOSE TO PROVIDE INFORMATION ON THE OPERATIONS BRANCH TO ASSIST IN FUTURE PLANNING BY:-

a. VERIFYING AND QUANTIFYING ALL TASKS BEING PERFORMED BY THESE RATINGS

b. CREATING A DATA BANK OF INFORMATION FROM WHICH JOB SPECIFICATIONS AND JOB COMPARISONS CAN BE DRAWN

c. OBTAINING OPINIONS ON CERTAIN FACTORS RELATING TO JOB SATISFACTION

d. COLLECTING DATA ON VARIOUS ASPECTS OF SERVICE CONDITIONS

OPERATIONS BRANCH SURVEY

EXTRACT FROM OPERATIONS BRANCH TO NMUU DIRECTIVE

SECONDARY PURPOSE WHERE POSSIBLE, DURING THE SURVEY, INFORMATION IS TO BE OBTAINED ON THE FOLLOWING SUBJECTS WITH REGARD TO OPERATIONS BRANCH RATINGS:-

a. SEAMANSHIP DUTIES
c. COMMUNAL DUTIES
e. COMMON OPERATOR TASKS
g. ADMINISTRATION OF ON JOB TRAINING
k. EMPLOYMENT OF LEADING RATES, SENIOR RATES, ACTING, LOCAL ACTING AND PASSED FOR HIGHER RATE PERSONNEL, AND TRAINING RECEIVED
l. ATTITUDES TO ADVANCEMENT
n. THE EMPLOYMENT OF THE OPERATIONS BRANCH CO-ORDINATOR.

OPERATIONS BRANCH SURVEY STATISTICS

	POPULATION	SAMPLE
GENERAL SERVICE	10,500	7,700 (73%)
SUBMARINE SERVICE	1160	820 (71%)
GENERAL SERVICE SUPERVISORS		500
SUBMARINE SERVICE SUPERVISORS		150

SLIDE 7

OPERATIONS BRANCH PILOT SURVEY

BACKGROUND INFORMATION

BILLET

QUALIFICATIONS

SHIP/EMPLOYMENT

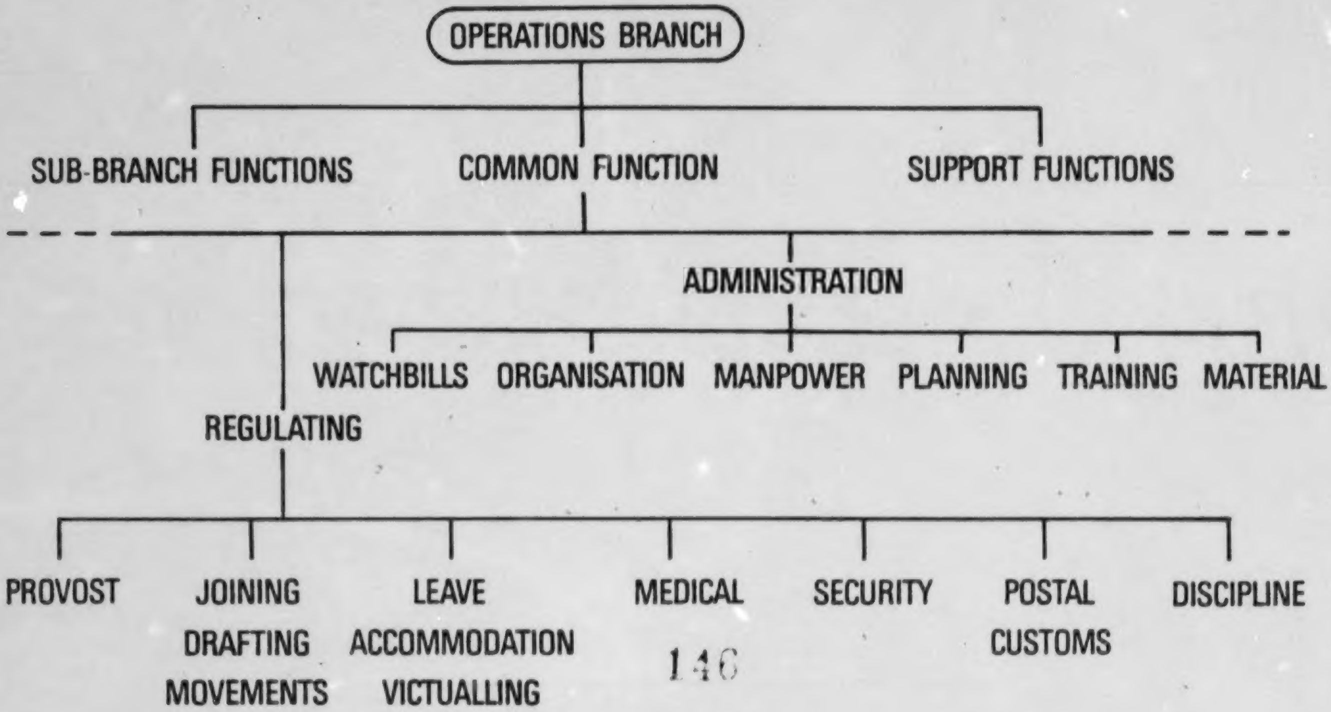
PRIMARY/SECONDARY WORK AREA

EMPLOYMENT AT ACTION STATES/CONDITIONS

SEAMANSHIP TOPICS

OPERATIONS BRANCH SURVEY

SCALAR DIAGRAM EXAMPLE



OPERATIONS BRANCH

SUB-BRANCH FUNCTIONS

COMMON FUNCTION

SUPPORT FUNCTIONS

ADMINISTRATION

WATCHBILLS

ORGANISATION

MANPOWER

PLANNING

TRAINING

MATERIAL

REGULATING

PROVOST

JOINING
DRAFTING
MOVEMENTS

LEAVE
ACCOMMODATION
VICTUALLING

MEDICAL

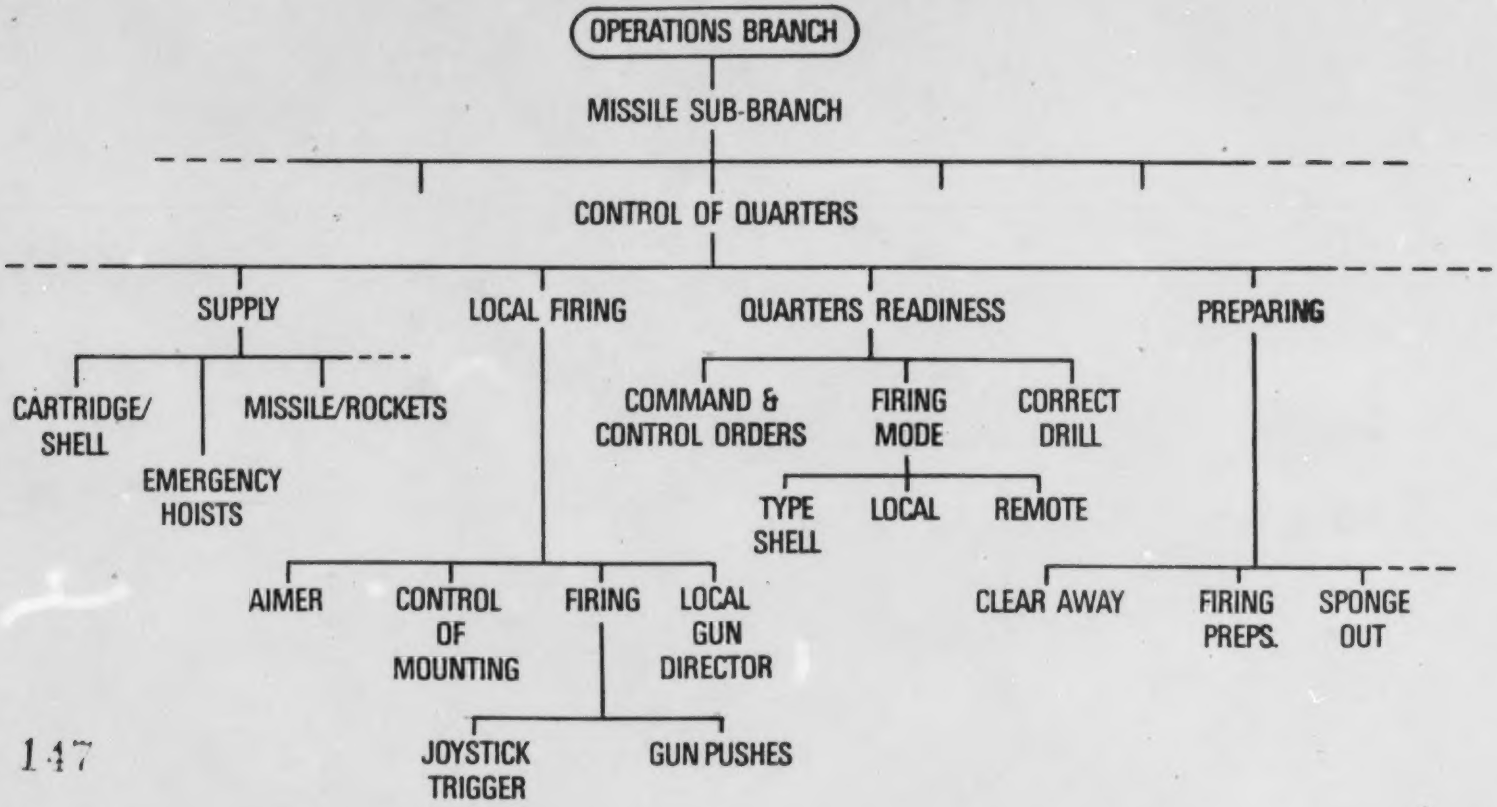
SECURITY

POSTAL
CUSTOMS

DISCIPLINE

OPERATIONS BRANCH SURVEY

SCALAR DIAGRAM EXAMPLE



- OPERATIONS BRANCH
  - MISSILE SUB-BRANCH
    - CONTROL OF QUARTERS
      - SUPPLY
        - CARTRIDGE/ SHELL
        - EMERGENCY HOISTS
        - MISSILE/ROCKETS
      - LOCAL FIRING
        - AIMER
        - CONTROL OF MOUNTING
        - FIRING
          - JOYSTICK TRIGGER
          - GUN PUSHES
        - LOCAL GUN DIRECTOR
      - QUARTERS READINESS
        - COMMAND & CONTROL ORDERS
        - FIRING MODE
          - TYPE SHELL
          - LOCAL
          - REMOTE
        - CORRECT DRILL
      - PREPARING
        - CLEAR AWAY
        - FIRING PREPS.
        - SPONGE OUT

THANK YOU FOR YOUR

NMUU
GET
STUF

CO-OPERATION

SLIDE 10

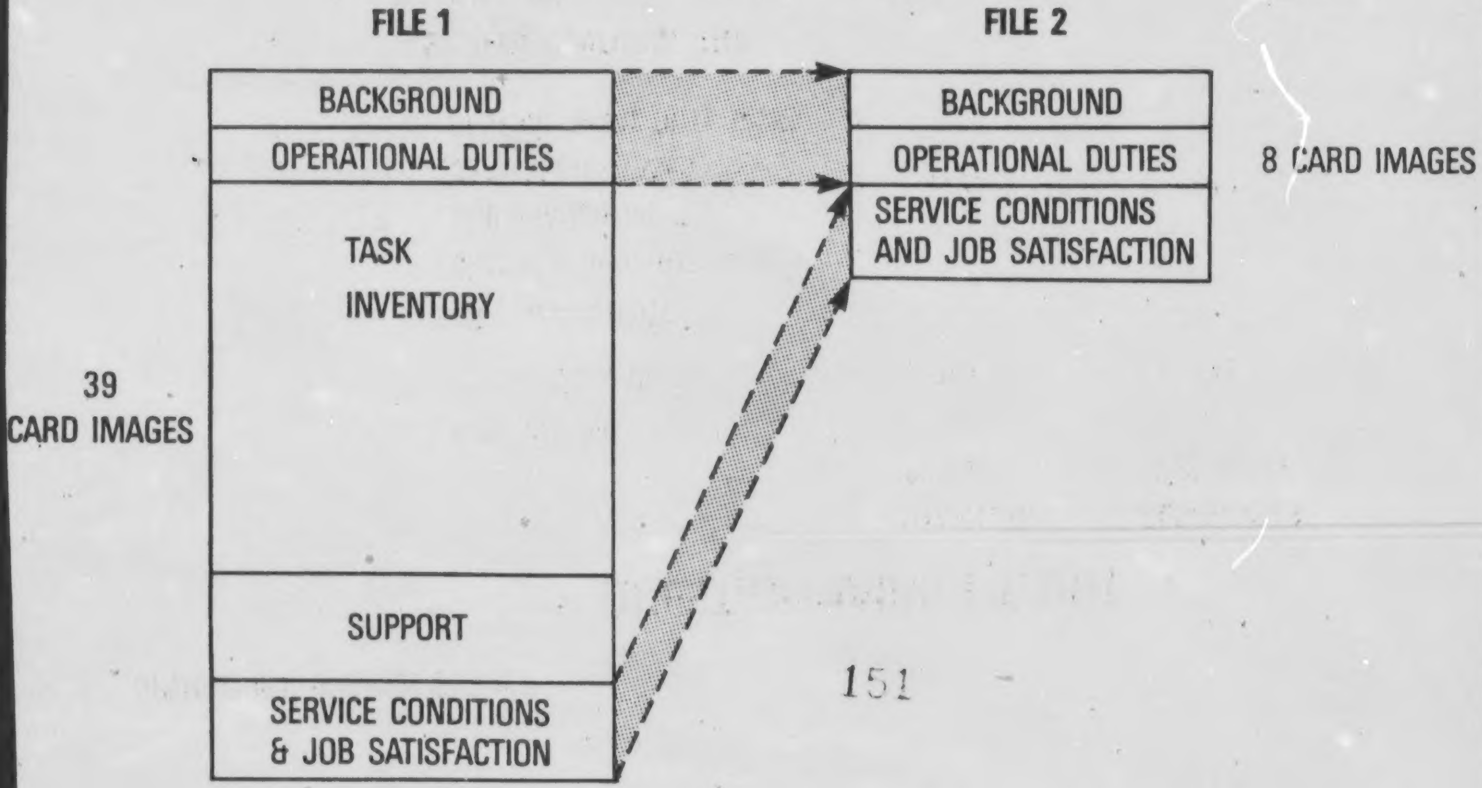
OPERATIONS BRANCH SURVEY

QUESTIONNAIRE LAYOUT

	POTENTIAL	RESPONSES
	PRIMARY	SECONDARY
BACKGROUND	31	
OPERATIONAL DUTIES	261	142
TASK INVENTORY		
COMMON OPERATOR TASKS	1137	848
SUB-BRANCH TASKS		
SEAMANSHIP TASKS		
GENERAL NAVAL DUTY TASKS		
SUPPORT INFORMATION		
COMMON	431	
SPECIALIST		
GENERAL NAVAL DUTIES		
SERVICE CONDITIONS	81	
JOB SATISFACTION	50	

OPERATIONS BRANCH SURVEY

COMPUTER FILE CONSTRUCTION



FILE 1

BACKGROUND

OPERATIONAL DUTIES

TASK INVENTORY

SUPPORT

SERVICE CONDITIONS & JOB SATISFACTION

39 CARD IMAGES

FILE 2

BACKGROUND

OPERATIONAL DUTIES

SERVICE CONDITIONS AND JOB SATISFACTION

8 CARD IMAGES

OPERATIONS BRANCH SURVEY

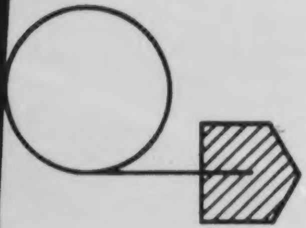
PUBLICITY & P.R.

1. ARTICLES IN "NAVY NEWS"
2. DEFENCE COUNCIL INSTRUCTION
3. LETTERS TO ALL SHIP/ESTABLISHMENT COMMANDING OFFICERS
4. APPOINTING OF SURVEY LIAISON PERSONNEL IN SHIPS AND ESTABLISHMENTS
5. QUESTIONNAIRE FOREWORD BY VICE ADMIRAL (DGNMT)
6. UK SHIP/ESTABLISHMENT VISITS BY NMUU PERSONNEL

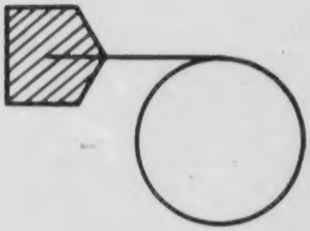
SLIDE 17

OPERATIONS BRANCH SURVEY

DATA FILE CONSTRUCTION



1. TAPE TRANSLATION
2. CHECK OF MANUAL CODING
3. SECONDARY CODING FOR SHIP/ESTABLISHMENT CLASSIFICATION
4. SORTING
    a. BY SUB BRANCH
    b. BY RATE
5. MERGE INTO MASTER FILE



OPERATIONS BRANCH SURVEY

REPORTS OF ANALYSIS – EXAMPLES

1. **JOB DESCRIPTIONS**
   EMPLOYMENT OF SENIOR RATINGS.
   RADAR SUB-BRANCH TASKS AND DUTIES.
   COXSWAINS OF BOATS.
2. **SERVICE CONDITIONS**
   ATTITUDES TO ADVANCEMENT.
   LIVING CONDITIONS IN SHIPS.
   BANK ACCOUNTS.
3. **TRAINING**
   GENERAL REPORT ON TRAINING ADEQUACY.
   RADAR SUB-BRANCH TRAINING.
4. **ORGANISATION**
   ADMINISTRATION, COMMUNAL AND GENERAL NAVAL DUTIES.
   SHIP HUSBANDRY AND CLEANING.
   CEREMONIAL.

A STRATEGY FOR TASK ANALYSIS AND CRITERION DEFINITION BASED ON MULTIDIMENSIONAL SCALING

The views and opinions expressed in this paper are those of the author and should not be construed as the official position of the Department of National Defence

Paper prepared for the:
Twentieth Annual Conference of
The Military Testing Association
Oklahoma City, Oklahoma
October 30 - November 3, 1978

155

Lieutenant-Colonel
Glenn M. Rampton
Commanding Officer
Canadian Forces Personnel
Applied Research Unit

Abstract

While the technical literature pertaining to independent measures (such as aptitude tests, vocational interest inventories, and so on) is burgeoning, much less time and attention is paid to advancing the technology of dependent or criterion measures. One reason that useful approaches for handling the "criterion problem" have been slow to evolve is that procedures required to surmount certain technical aspects of the problem have yet to be developed, or are not widely known. Another reason is that, although relevant techniques for handling other aspects of the problem have been published, insufficient systematic effort has been expended to integrate them into practical research strategies. A research strategy using nonmetric multidimensional scaling was developed to fill in some of these practical technological gaps. This was tested on Air Observers (operators of complex sensor and communications equipment used in antisubmarine and Northern surveillance aircraft in the Canadian Forces) The content dimensions produced in this application proved: (a) highly reliable and internally consistent within relatively homogeneous groups of individuals; (b) readily and meaningfully generalizable across a variety of work situations (responsibility levels); (c) valid in terms of showing significant relationship to external variables, and being readily integrated into larger bodies of scientific knowledge; and, (d) extendible in theoretically and practically important ways in other studies. A more comprehensive treatment of the results, discussion, and conclusions deriving from this research programme is available on request from the author. The present paper focuses specifically on the design and analytic methods used, since it is believed that, as a general research strategy, they have relevance for those involved in task analysis and criterion definition, particularly in human factors engineering, test and training validation, and performance evaluation applications.

In theoretical and applied psychological research one is often faced with: (a) defining what dependent or criterion measures are likely to be important in specific content areas, and (b) developing procedures to collect reliable, valid data reflecting these dimensions once they have been defined. However, while the technical literature pertaining to independent measures (such as physiological indicators; both written and other expressions of aptitudes, vocational interests, personality, attitudes, or job performance, and so on), is burgeoning, less attention has been devoted to advancing the technology of dependent or criterion measures. This is true of research having to do with selection and classification, training and education, performance evaluation, human factors, human and organizational development, and other areas of psychology where a sound knowledge of the performance content domain with which one is dealing should be the basic starting point for subsequent research. As Christensen and Mills (1967) point out:

> The criterion problem is much like the weather - all psychologists talk about it but very few do much about it. And yet its central importance is disputed by no one. Over twenty years ago Thorndike (Note 1, p. 29) attested to its importance in military operations when he said, "Certainly the most fundamental and probably also the most difficult problem in the Aviation Psychology Program was that of obtaining satisfactory criterion measures against which to validate tests and evaluate variations of training methods". (p. 335).

A number of papers have recently been published about various aspects of the "criterion problem" (e.g., Christensen & Mills, 1967; Dunnette, 1963; Inn, Hulin & Tucker, 1972; Crooks, (Note 2). This work has, as yet, failed to produce many concrete solutions. The discussions have generally been more useful in defining various aspects of the problem than in demonstrating how they might be handled.

In terms of what task analysis should be, or what it should accomplish, Miller (1963) has argued that task analysis should involve the systematic study of the behavioural requirements of tasks. Gagne (1963) has suggested that it should allow inferences based on the knowledge of human functions concerning what kinds of abilities, skills and knowledge are required in order for a human being to carry out specific tasks. Kershner (Note 3) has indicated that job analysis should answer the "what", "how" and "why" of tasks.

Ammerman (Note 4), on the other hand, has been a bit more explicit, suggesting that task analysis should: (a) yield an organizational scheme accounting for all previous knowledge of relevant job activities, (b) identify and account for all activities relevant to the specific job, (c) take into account and be consistent with psychological concepts of human behaviour, and hopefully, be generalizable to a range of jobs, and (c) be temporarily practical and meaningful to users. Dumas and Muthard (Note 5) have argued that task analysis should also offer: (a) classification of tasks in situ to minimize the introduction of errors, (b) measurements that are reliable and on interval scales insofar as is possible, and (c) a methodology which is compatible with appropriate system analytic and operations research techniques so that critical decisions made about specific aspects of the job can be simultaneously related to other important data elements.

It is difficult to argue with these lists of desirable task analysis characteristics and objectives. In a sense they have the glow of motherhood. Unfortunately, in themselves, they do not imply how these ends are to be achieved. This fact notwithstanding, the points raised were regarded as desirable goals for the task analysis procedures outlined in succeeding sections.

Task Description Versus Task Analysis

Most conventional task analysis strategies have been limited to the use of specific data gathering procedures in conjunction with a rational taxonomy. The intent in these studies is to classify task elements according to psychological constructs reflecting the particular theoretical predilections of the investigators.

Breaking a job down into a number of reasonably elementary components and then rationally classifying these according to some scheme can be useful as a first step in a larger program. This process does not go much beyond what Miller (1963) calls "task description". In addition to these preliminary data collection and organizing phases, one requires means for obtaining a behavioural understanding of the task requirements. Miller has reserved the term "task analysis" for this latter process.

Fleishman (1967a, 1967c) and Finley, Obermayer, Bertone, Meister and Muckler (Note 6) have argued that investigators must strive to move beyond the mere identification and classification of discrete task elements in specific work settings to the distillation of a relatively parsimonious set of unifying fundamental behavioural elements gathered from a number of settings. The Fleishman approach has tended to involve examination of various aspects of performance in laboratory settings. This has produced some very useful information but cannot avoid suffering from a certain amount of artificiality since it ignores (and in many ways is designed to eliminate) contextual factors. The importance of the context in which tasks are performed is well recognized (see discussions by Alluisi, 1969; Christensen & Mills, 1967; Grodsky, 1967; Miller, 1963; Prien & Ronan, 1971), and any deemphasis of it could not help but constitute a major weakness of this approach. Finley et al. (Note 6) have argued for the identification of "fundamental behavioural dimensions" underlying tasks identified in the "man-machine" environment, but after conducting a fairly comprehensive review of the literature were unable to suggest how this might be done.

Multidimensional Scaling

It was felt that multidimensional scaling, in conjunction with regression analysis, and allied multivariate techniques might be suited to the kinds of analyses called for in the preceeding section. In general, given a matrix of numbers showing how similar each object in a set is to each of the remaining objects in the same set, the goal of multidimensional scaling (MDS) is to determine the minimum dimensionality of the relationships as well as the projections or scale values of the objects on each of the resulting dimensions. This, of course, is precisely what one would like to do in task analysis.

MDSCAL (Kruskal, 1964a, 1964b) and other nonmetric multidimensional scaling algorithms require input in the form of stimulus by stimulus similarity (proximity) matrices showing but ordinal interrelationships among the stimulus objects under study. For these data, the algorithms attempt to derive a representation of n points (representing the objects) in a geometric space of smallest dimensionality such that the original proximities (let these be represented by B_{ij}), and the final geometric interpoint distances (let these be represented by d_{ij}) are related monotonically. That is, so the geometric interpoint distances $d_{ij} < d_{kl}$ when the similarities $B_{ij} > B_{kl}$ (if the Bs are dissimilarities, one requires $B_{ij} < B_{kl}$).

The analysis proceeds through a series of successive iterations. One starts with an arbitrary initial configuration (of known dimensionality in n points) which may be randomly generated; a "best guess" on the part of the investigator, or created in a number of other ways (e.g., the Young/Torgerson option used in the computer programme KYST - see Kruskal, Young & Seery, Note 7). Starting from this initial

159

configuration the n points are adjusted mathematically such that, in a space of specified dimensionality, their distance interrelationships (d_{ij}) more and more closely reflect the monotonic (ascending or descending) interrelationships of the respective B_{ij}. The procedure continues until one of a number of criteria has been met which indicates (either absolutely, or in a practical sense) that no more improvement in the solution is possible. The values of these criteria may be specified by the investigator, and relate to the number of iterations conducted, how fast the solution is converging, or how well the monotonic d_{ij} vs B_{ij} requirement is met.

The index of how well the monotone relationship between the B_{ij}s and d_{ij}s is met in a particular iteration has been referred to as stress (Kruskal, 1964a). A zero stress value indicates that a perfect monotone relationship exists between the dissimilarities and final fitted d_{ij}s. Rampton (Note 8) presents a more extensive conceptual discussion of this analytic model. One generally conducts separate analyses on the same data, in a number of dimensionalities. One then chooses among the separate solutions on the basis of goodness of fit (low stress), parsimony (adequate representation in fewest dimensions), and interpretability (the solution should make sense).

One might question the appropriateness of using data interrelationships reflecting only ordinal qualities of measurement to generate a metric configuration. In discussing the rationale underlying nonmetric MDS, Shepard (1962) has argued that knowledge of ordinal relationships of distances really implies much stronger than ordinal measurement when the points to which the distances refer are considered in the context of a configuration of known dimensionality. Further, the greater the ratio of numbers of points to numbers of dimensions, the more finitely the final configuration can be determined.

Method

Participants

Participants in this research project were either members of the Air Observer trade in the Canadian Forces, or individuals having an intimate working knowledge of it. The Air Observer is the primary operator of sophisticated sensing and communications systems on military ocean and Northern surveillance aircraft. To gain entry to this trade, an individual must have been trained and have a good record in another trade in the Canadian Forces. He must also have achieved a minimum standard on a test of general learning ability. Then, if selected in competition with others meeting these criteria, the individual undergoes a demanding programme of aircrew training. On the job itself, the Air Observer must remain vigilant while monitoring equipment over long periods of time. These periods are interspersed with sessions of rather intense and critical activity.

Specific samples in this research programme included: (a) two groups each of 28 Air Observers (henceforth referred to as OBS1 and OBS2), (b) 21 Air Observers (referred to as SUPS) holding supervisory positions, (c) five students (referred to as STUDS) undergoing final stages of qualification training, and (d) eight commissioned officers (referred to as ROS) with extensive experience in operations, operational training, and staff capacities associated with the trade.

Task Definition

As a first step in defining the content domain for further study, training manuals and checklists covering the range of Air Observer duties in the Argus long range patrol aircraft were formulated into task elements and classified according to the Berliner, Angell and Shearer (Note 9) taxonomy.

As a cross-check, the task elements were reviewed for completeness and independently categorized according to the taxonomy by instructors at the Maritime Operational Aircrew Training Squadron, Canadian Forces Base Greenwood, Nova Scotia. These individuals were well acquainted with all aspects of the Air Observer job, since each had many hundreds of hours experience with it, both in training, and in operational capacities. The information from this step was compared to that from the former one.

In a third iteration of the procedure, the task elements generated in the former two steps were categorized according to the taxonomy by senior operational personnel at Canadian Forces Base Greenwood. These results were compared to the composite of the former two steps. In each of the separate applications of the taxonomy, when discrepancies were found, these were resolved by negotiation with representatives of the various groups involved. In most cases, consensus was easily achieved.

Finally, the author and two colleagues flew several operational training missions with Argus crews. The purposes of these flights were to offer an intuitive idea of some of the contextual and environmental circumstances in which the tasks are performed.

The procedures described above produced more than 350 task elements, which were far too many to be handled by existing MDS strategies. Further, the elements differed in level of abstraction (a list of the elements is presented in Rampton, Note 8). In an attempt to come to grips with these problems, the total list was reviewed with the help of an officer with extensive operational experience (more than 2500 hours in the Argus aircraft), and reformulated into 166 task functions. These functions were generated so that all were at about the same level of abstraction, and were couched in phraseology and jargon that would be readily understood by the Air Observer. Three statements were added to this list on the basis of pilot work with the experimental procedures. The resulting list of 169 task functions is presented in Rampton (Note 8).

Materials

The materials assembled for each participant involved:

1. One or two decks of computer cards, each containing 169 cards on the top of which were printed the 169 task functions (one to a card). Each card also had a unique identification code punched, but not printed in columns 72-80.

2. One or two white computer cards on which were printed spaces for identification and other pertinent information.

3. One or two decks of 16 blue pile-separator cards each containing one of the numbers from 1 - 16.

4. One or two sheets of paper on which all possible pairings of the numbers from 1 - 15 (a number was not paired with itself) were arranged in random order, making n(n-1)/2 = 105 pairs.

5. A booklet containing all task statements. Accompanying each of these booklets were three sets (of a possible seven) of five point scales. The

full set of variables were: (a) degree of concentration required in performing the task, (b) difficulty level of the task, (c) manual skill required in performing the task, (d) the importance of the task to successful completion of missions in which it is typically performed, (e) the cooperation or teamwork generally required to complete the task, (f) the importance of speed or working quickly to successful completion of the task, and (g) the degree of mental effort (decision making, calculating, memory, and planning) required to successfully complete the task.

The duplicate materials alluded to in points 1 - 5 above, were used in a test - retest reliability study of the sorting task described below.

Experimental Design

Pilot work had suggested that order effects might be important in the presentation of the task statements and answer sheets. Therefore, as a first step in incorporating a partial balance into the presentation of the task statements to participants, the following four blocks of statements were created: (a) thirty items relating to antisubmarine warfare, (b) fifty items relating to electronic counter-measures and communications, (c) forty-nine items relating to detection functions, and (d) forty items relating to the use of RADAR. These blocks were independently organized into two four-by-four Latin squares. One of these was used for balancing the presentation of items in the card deck, and the other was used for presenting the items in the task statement booklet. To control for order effects in presentation of the three rating scales to be used by each person (time constraints dictated that all seven sets of scales could not be done by each), these were arranged in a Youden square design.

The design precautions outlined above provided four different task booklet combinations, four different card deck combinations, and seven different answer sheet combinations. As an additional control for any interaction between book type and answer sheet presentation, the book and answer sheet combinations were arranged in blocks of twenty-eight (i.e., 4 books x 7 answer sheet combinations = 28) so that each answer sheet combination was paired with each of the book types. Within each block of twenty-eight book/answer sheet combinations, the four deck types were assigned so that seven of each type were randomly represented in each block.

Integers from one to twenty-eight were then randomly assigned to the book/answer sheet/deck combinations. These numbers represented the order of presentation of specific treatment combinations to participants. Four separate blocks were created in this way with assignment of participants to treatments being independently randomly assigned in each. The first two blocks of this design were reserved for the OBS1 and OBS2, while the third block was reserved for the first twenty-eight SUPS and ROS. The remaining participants (the STUDS and others) were assigned in order to treatment combinations in the fourth block.

Procedure

The experimental procedure followed is outlined in more detail in Rampton (Note 8), but basically consisted of having participants sort, and then subsort the piles of task statements on the basis of the similarity they felt existed in the performance of the functions on each of the cards. Free, but not completely unconstrained sorting was used, in that a maximum of 15 piles were to be used in the major sorts (with an additional "miscellaneous" pile to be created only if absolutely necessary), with a maximum of five to be used in the subsorts (again, with an additional sixth pile was to be used if required). The maximum numbers of piles for both the sorts and subsorts were chosen on the basis of what a rather extensive pilot project suggested were more than required by most participants.

Between the sort and subsorting stages, participants were asked to serially number all of the major piles which they had placed on the table in front of them. On a sheet of paper containing 105 scales each with the headings "Category X", "Category Y", (where X and Y stood for the numbers attached to the categories) and the numbers from 1 to 5, participants rated the similarity of all possible pairings of the constructs reflected in each of the piles.

After the above steps were completed, a booklet containing all task statements was distributed to each individual. Inserted inside the back cover of the booklet were three sheets of defined scales to allow rating of each task function on the variables: (a) Concentration Required, (b) Difficulty, (c) Manual Skill Required, (d) Importance, (e) Cooperation or Teamwork Involved, (f) Speed Required, and (g) Mental Effort Required. Taking each of the three sheets of paper separately, participants were instructed to go through the task statement booklet three times, and to rate each task according to the variables defined on the respective sheets of paper.

After all data were collected, a card containing biographical information was keypunched and concatenated with the pile rating and other data described above to form the test data set for each person. A computer programme assigned proximity indices between task statements for each individual as follows: (a) if two statements were not grouped together they received a proximity index of 6 - X; where X was the similarity rating assigned their respective major categories, (b) if two statements were grouped together after the first sort but not after the second, they received a conjoint score of 7, and (c) if two statements were grouped together after the final sort they were assigned a score of 9. Stimulus by stimulus half matrices (without diagonal) having (169 x (169 - 1) 1/2 = 14,196 similarity estimates as entries were thus produced for each individual.

Analyses and Brief Discussion of Results

Reliability of the Sorted Indices

Thirty-one of the journeyman (OBS1 and OBS2) Observers repeated the sorting and pile rating stages for a test-retest reliability study. Each individual received the same combinations of materials in both sessions except that on retest, the unidimensional ratings were not done.

A more comprehensive justification for the use of sort-generated proximity indices as input for multidimensional scaling, their reliability, and validity, is presented elsewhere (Rampton, Note 8). Suffice to mention here that dissimilarity indices produced by taking arithmetic means across individuals in each of the test and retest sessions, and then computing a Pearson *r* down the respective aggregate proximity indices between sessions, produced a correlation of .94, thus indicating considerable retest reliability.

Thirty of the individuals who participated in the test-retest reliability study were divided into groups on an even-odd basis. Two aggregate proximity matrices were created by taking arithmetic means of the dissimilarity indices across individuals in each group. A correlation was computed down the respective aggregate proximity indices in the test and retest sessions producing within-group consistency correlations of .83 and .83, respectively. These values give further evidence that considerable consistency existed in the way that different individuals perceived the task statements.

Nonlinear MDS Analyses

Five half matrices consisting of average similarity indices for the ROS, SUPS, OBS1, OBS2, and STUDS were calculated by computing arithmetic means of respective stimulus by stimulus values across all individuals in each group. An additional "total average" (AVE) matrix was created by calculating analogous indices over all participants. Data in these six matrices served as the basic input for the MDS analyses.

As nonmetric MDS programmes iterate toward a goal of stress minimization, they may get caught up in less than optimal solutions by locating local function minimums. This is more likely to occur the more dissimilar the initial configuration is from the "optimal" configuration. Spence (1972), in a rather extensive empirical comparison of a number of MDS strategies, indicated that a procedure developed by Young and Torgerson (1967) may effectively circumvent local minimum problems. This algorithm, which involves using conventional metric MDS on input data to produce an initial configuration for the nonlinear MDS was modified slightly and used to start MDSCAL analyses of the AVE proximity matrix. As shown in steps 1 to 3 of the schematic analysis representation of Table 1, this modification involved creating a randomly augmented "initial" matrix for MDSCAL of the AVE data, and was required since KYST can handle proximity matrices reflecting but 60 elements vice the 169 involved in this project. Figure 1 represents a plot of the stress values of AVE MDSCAL analyses in configuration dimensionalities from ten down to two.

Insert Table 1 and Figure 1 about here

The resulting 169 x 10 AVE configuration was used as the initial configuration for MDSCAL analyses of each of the ROS, SUPS, OBS1, OBS2, and STUDS proximity data. Numbers of iterations and stress values for these analyses are listed at the bottom of Table 1.

Representation of the task dimension scale values would require more space than is justified here, but may be obtained from Rampton (Note 8). Visual inspection of these values showed considerable similarity across configurations.

Comparing MDS Structures Using Regression Analysis

Multiple linear regression (REGRESS - see Miller, Shephard & Chang, 1972, for a discussion of the specific technique used) provided the means for a more rigorous comparison across configurations. Results of these analyses are presented in Tables 2 to 5. In the terminology of the traditional test - criterion validation paradigm: (a) the ten dimensions of the SUPS, OBS1, OBS2, and STUDS configurations served as "predictors" in separate analyses, (b) each of the 10 ROS dimensions served in turn as a "criterion" in each set of analyses (SUPS vs. ROS, OBS1 vs. ROS, OBS2 vs. ROS, and STUDS vs. ROS), and (c) the 169 task statements served as "subjects" in each run. One can conceive of these analyses as equivalent to locating directions or vectors in the SUPS, OBS1, OBS2, and STUDS configurations correlating most highly with each of the 10 ROS dimensions. Thus the multiple correlations (Rs) of Tables 2 to 5 reflect the strength of relationship between these best fitting, artificial ROS vectors, and the actual ROS dimensions. The Rs are seen to be generally quite large. (Computing confidence intervals in the manner suggested by Garrett (1966, p. 416) indicates that a critical value of $R \geq .20$ is required to be statistically significant at $p < .01$ for each of the multiple Rs reported in this section).

Insert Tables 2 to 5 about here

The ROS were chosen as the primary reference group in these analyses because: (a) they generally had more experience with the tasks than did individuals in the other groups, and (b) they were all senior supervisors, trainers, or responsible for maintaining proficiency/performance standards. For present purposes, this latter point is particularly relevant. The three functions subsumed within it imply that these individuals should tend to represent what might be called the "official point of view" about technical aspects of task performance.

The matrix of direction cosines in each of the tables shows that these fitted ROS vectors and the initial configuration dimensions for each group corresponded in a one to one fashion. While the SUPS, OBS1, and OBS2 configurations differed little in the degree to which they related to the ROS dimensions, the STUDS data did not show as much correspondence.

Fleishman (1967b) has shown that as people become more proficient at complex tasks, different kinds of abilities contribute to performance. This might explain why the STUDS' configuration did not show as much correspondence to that of the ROS as did those of the more experienced groups. Fleishman's observation would also lead one to expect that those groups most alike in experience and proficiency would perceive underlying dimensions of their jobs more similarly than groups less alike on these variables. To test whether this might be borne out in the present data, Rs were calculated between dimensions from the OBS1 configuration and each of the OBS2 dimensions. Table 6 summarizes the results of these analyses.

Insert Table 6 about here

These two configurations can be seen to correspond very highly, both in content and orientation.

Cosines of .00 existed between the original MDSCAL ROS dimensions. In Tables 2 to 5 however, one observes that many of the cosines (correlations) between the ROS vectors, when fitted into the configurations of remaining groups are considerably larger than .00 in absolute value. (A cosine or correlation of .00 denotes an angle of 90°). For example, eight, nine, six, and twelve ROS intervector cosines in the SUPS, OBS1, OBS2, and STUDS configurations, respectively, exceeded .30 in absolute value. Seven of the cosines in the STUDS configuration had more than three values this large. Further, the two largest values in the STUDS configuration (.62 and - .75) far exceeded the next largest values in any of the other configurations. This evidence strongly suggests that the groups responded to the task functions in systematically different ways. It also indicates that the STUDS differed more from the ROS in this regard than did the other groups.

Contrary to expectation, however, the SUPS did not appear to be significantly more like the ROS than did the OBS1 and OBS2. (A probable reason for this is given in Rampton (Note 8), and relates to historical training and experience commonalities shared by the OBS1, OBS2 and ROS that were not as similar for many of the SUPS). Smaller intervector correlations did result when the OBS2 content dimensions were inserted into the OBS1 configuration (as shown in Table 6, only five of the OBS2 fitted vectors in the OBS1 configuration exceeded .30). This indicates that the OBS1 and OBS2 formed a relatively homogeneous dyad when considered in the context of the four groups of "skilled" participants. In total, this evidence is taken as supporting a contention that the more similar two groups are in skill, experience level, and other characteristics, the more similarly they will perceive salient aspects of their work.

Relation Between Configurations and Rated Properties

Average unidimensional ratings for each of the 169 task statements of the seven variables defined earlier (see Rampton, Note 8, for the instructions and format under which these scales were administered), were calculated by taking arithmetic means of respective task ratings over all individuals. Numbers of respondents per scale were: Concentration (37), Difficulty (38), Manual Skill (41), Importance (37), Cooperation (39), Speed (39), and Mental Effort (36). These numbers were not all equal because: (a) the scales, Importance, Concentration, and Mental Effort, for one of the OBS2 individuals were not completed properly and had to be discarded, and (b) the fourth Youden square was incomplete (containing but six participants), so that the answer sheet balance inherent in each complete block was unfulfilled in the last one.

REGRESS was used to locate vector orientations in the AVE, ROS, SUPS, OBS1, OBS2, and STUDS configurations showing maximum correspondence to each of the average rated properties. The results of these analyses are shown in Tables 7 to 12. The format of these tables parallels that of Tables 2 to 6. Multiple correlations or Rs between the 10 dimensions for each group and each of the seven rated properties are shown as the first line of numbers in each table. Each table also provides a matrix of direction cosines showing how the fitted vectors were oriented in the respective configural spaces, as well as a matrix of cosines showing the interrelationships of the vectors in the space.

Tables 13 to 22 represent an attempt to interpret each of the ten AVE content dimensions produced in the analyses. Each table contains a statement summarizing the definitions derived for each dimension. Each also contains a listing of twenty of the more salient tasks (ten on each end of the dimension) to serve as typical representatives of these constructs. Interpreting these dimensions turned out to be a complex process. While the data in Tables 7 to 12 were the primary sources of information used, simultaneous consideration of this information with virtually all that contained in Tables 2 to 6, and the loadings of the tasks on the dimensions were necessary. It was quickly discovered that an intimate working knowledge of each of the tasks was also essential, and the author was fortunate in being able to rely on colleagues at the Canadian Forces Personnel Applied Research Unit, (having considerable experience with the content domain under study) and experienced Navigator/Radio Officers working at military establishments in the local area to assist him in this regard.

While the Rs in Tables 7 to 12 are all significant at $P<.01$, they are generally small to moderate in absolute value. (This is perhaps understandable given the inherent limitations of this kind of criterion measure). Further, one notices similar multiple correlation and direction cosine profiles across tables. Difficulty and Manual Skill tend to have lowest saliences in each configuration; Importance, Cooperation, and Mental Effort generally have moderate salience; while Speed and Cooperation typically show largest relationship.

The matrices of direction cosines showing correspondence between the fitted vectors and the configural dimensions, as well as the matrices of cosines of angles between property vectors in the configurations were useful for interpretive purposes. They depict the relationships among the properties and dimensions as well as the interrelationships between the properties when located in the configuration.

In examining the evidence in Tables 7 - 12, it is important to remember that though the Youden square arrangement was set up to balance presentation of the scales, this balance was not complete since the last experimental block was only partially filled. One should also be aware that the properties are somewhat confounded. This granted, it is apparent from the matrices of cosines of angles between the property vectors in the configurations, that the property ratings reflected more than subject variance confounding, or halo. Further, the relationship profiles are reasonably consistent across all configurations, and make a great deal of intuitive sense. For example, the properties Concentration, Difficulty, Importance, Speed, and Mental Effort show moderate to large interrelationships. The only variable which shows consistent relationship to Cooperation is Manual Skill, reflecting the fact that many of the heavy, physical tasks done on the Argus aircraft by an Air Observer are typically done in cooperation with someone else.

Implications and Possible Research Extensions

The results of Tables 2 to 12 make sense when considered solely on the basis of the structural representations, as well as when considered in the context of external criteria. This augurs well for the validity of the form and content of the dimensions produced, and thus the methodology used to produce them. However, some question might still remain as to the relevance of these data in the context of the "criterion problem". One might, for example, question whether the task functions as derived were

170

the most useful entities on which to base the dimensional analyses, and whether actual or simulated job problems (such as the circuit types typically repaired by naval aviation technicians used by Schultz & Siegel, Note 10, or the simulated air traffic control situations used by Landis, Silver, Jones & Messick, 1967) might not be more appropriate. These and a number of related issues are discussed in succeeding paragraphs.

Appropriateness of Task Functions as Basic Analytic Units

The Study of Operator vs. Task Characteristics. The basis on which similarity of dissimilarity judgements are made in a MDS study must have an important bearing on results. Both the Schultz and Siegel (1962) and the Landis et al. (1967) studies required participants to make proximity judgements on the basis of the similarity of the stimuli per se, while the Air Observers and Radio Officers were asked to make their decisions on the basis of how similar the tasks were to do. This difference in emphasis is believed important. In a sense, the distinction relates to the difference between the analysis of task performance in terms of operator characteristics versus task characteristics discussed by various authors (e.g., Prien & Ronan, 1971; Wheaton, Note 11). It seems obvious that dimensions produced from MDS of proximities based solely on judgements of similarity (or dissimilarity) of job problems or the like, must have a primarily "task characteristics" perspective, and inferences about ability/skill components will likely be possible only indirectly, through consideration of the configurations in the context of personal correlates (as was done by Landis et al.).

In the present investigation, an attempt was made to maintain an "operator" perspective. This was the reason for orienting instructions so that participants would respond on the basis of how similar the tasks were to do rather than on the basis of other attributes. Although analyses based on judgements of task similarity per se may be of interest in other applications, the outcome would not likely reflect the kinds of criterion (e.g., ability/skill) dimensions under investigation here. These comments should not, however, be construed as criticism of other approaches. For example, the purpose of the Landis et al. investigation differed from the one reported here. Thus, even if their approach were applied to the present situation, dissimilar (though hopefully complementary) results would be expected. A potentially useful extension to the research in this investigation, in fact, might be to prepare a number of tactical or other situations typically faced by Air Observers and to use these in place of the task functions in proximity generation and MDS procedures analagous to those outlined earlier. A comparison with the dimensions produced to those in this project could then be made. One would not expect the two sets of configurations to overlap completely, but one would hope that they

171

would be meaningfully relatable to each other.

Number and Specificity of Task Entities. In reviewing the literature for studies that have used MDS or related analytic approaches in task analysis, one is struck by the relatively small number of entities (tasks functions, simulated aerodrome situations, etc.) typically involved. For example, Brown (1967) used a sample of 18 task statements. Siegel and Schultz (Note 12) also used 18 task statements, Smith and Siegel (1967) used 34 task functions, and Landis et al. (1967) used 30 simulated air traffic control problems. In many of these applications, the investigators may have been limited either by the numbers of objects their computer programmes could handle, or by the amount of labour their method of generating proximities required of participants.

A number of procedures have been designed to economize on participant labour. Other procedures have been created which use separate computer runs to build up MDS solutions containing more objects than can be handled in a single run. Kruskal et al. (Note 7) provides a brief introduction to some of these techniques. Alternatively, if one has access to sufficient computing resources, it is sometimes possible to enlarge[1] the computer programmes to handle as many stimuli as are needed. Although these procedures have been available for some time, most investigators have either reduced the number of tasks by picking a small sample of all those possible, or defined the tasks at such a gross level of generality that a small number provided a global description of the job.

There are a number of potential difficulties in having relatively few task statements in a MDS analysis: with a small number of objects (and thus interrelationships) one cannot possibly obtain many dimensions, even if a larger number exists in the content domain. Further, the smaller the ratio, number of objects/number of dimensions, the less reliable or tightly constrained will be the final configuration.

In using MDS in task analysis, it is important to recognize that the level of generality with which the task statements are derived, though somewhat arbitrary, will have an important bearing on the results. For example, except for some work to make level of abstraction a bit more equitable, and to recast some in language that would be more readily understood by participants, the total set of 350 task elements might have served as the basis for similarity judgements and subsequent MDS analyses.

---

[1] Enlarging the MDSCAL and REGRESS programmes for this research project proved somewhat complicated. Copies of these computer programmes may be obtained from the author.

Vernon (1965) has proposed that one can view skilled behaviour as being structured hierarchically. At the apex are broad factors, each accounting for performance in a wide range of tasks. Below these, and serving as building blocks for them, are successive hierarchical layers of increasingly specific abilities, which, though pertaining to the same variance as the layers above, also account for some of the variance in more disparate tasks. From this perspective, the more specific and detailed one can be in generating task statements, the better. Thus, a comprehensive, specific list composed of many items should, other things being equal, account for more content variance than a general list composed of few. Within this conceptual framework, it should always be possible to produce a more general MDS configuration from a specific one by suitable rotation and/or clustering procedures, but not the converse--that is, one could not move from general solution to specific solution.

However, one is limited in the extent to which one can handle a long, detailed task statement list by the purely practical considerations of work capacity of participants, and computer resources. Even with a major effort to economize, both of these resources were stretched about as far as they could in this study. Thus the 169 task statements reflect a compromise between the desirability of having more and more detail vs. practical resourse constraints.

Generalizing Across Work Situations

The task analysis methodology used to investigate the Air Observer trade was designed to be as general as possible so that the same format could be used in many work situations. This was one of the major reasons for deciding on the two step task descriptive phase of: (a) breaking the job down into task elements and categorizing these according to an established task taxonomy, and (b) summarizing and rewording the content from the previous step into task functions designed to be at about the same level of generality and in language that could be understood by participants. With this process as a means of defining the content to study in each application, it should be possible to apply the methodology across trades with only minor adjustments.

In doing across-trade comparisons, one might start with a number of trades, each sharing content with at least some of the others. That is, one might have trade A sharing some components with trade B, trade B with trade C, A with D, and perhaps (but not necessarily) A with C. Note that each trade would not have to overlap with every other trade in the set, and that there need be no limit to the number of trades involved (except regarding practical constraints). One can imagine the situation as being represented by a Venn diagram of overlapping circles, each circle representing a trade. With the adding of more trades, it is possible that any two (say A and Z), though linked together by a pathway of other overlapping trades, might share no common variance.

Taking any two trades, for example A and B, one could process each through the task descriptive phase of the task analysis methodology, ensuring that the task functions derived for each were at about the same level of generality. Identically worded task functions corresponding to the content shared by the two trades would be generated and included in the total set of task functions for each trade.

Suppose that 100 and 109 task functions were created for trades A and B respectively, and of these, 41 were common. One would run through the proximity generation, MDS, and other methodological phases for each trade. Then, the 41 common task functions could be used as a nucleus around which to build a combined A + B configuration using the FIX option of an enlarged KYST computer programme (Kruskal et al., Note 7). One could repeat the procedure by adding trade C to obtain a combined A + B + C configuration, and/or separate A + C, B + C configurations. The FIX, KYST option, in conjunction with suitable algebraic manipulations, should allow one to infer interrelationships of tasks not shared by two jobs from knowledge of interrelationships of those that were. One would want to build in a number of cross checks to ensure that the results were consistent (i.e., in the configuration combination A + B + C one might start the process from different points -- e.g., C and B rather than A and B, to ensure the end result was the same).

In effect, the procedures as outlined should allow one to predict analytically, how tasks from different work situations might relate to each other if they were together, and could thus represent a powerful tool for the structuring and restructuring of jobs. Another important use of these procedures, of course, would be as a means of integrating task analysis data from complex work environments.

The MDS Dimensions as Criteria

Before the true significance of the methodology illustrated in this paper can be evaluated, it is necessary to establish how well, or indeed whether, the "potential" of the approach is translatable into reality. In the context of aptitude test development for example, one might ask whether any of the 10 performance dimensions were suggestive of kinds of tests that might predict training or job success in an applicant to the Air Observer trade. The following paragraphs outline research bearing on this point. It was conducted by a colleague of the author's at the Canadian Forces Personnel Applied Research Unit and is more fully documented elsewhere (Fournier, Note 18).

The Criterion Dimensions as a Basis for Developing Aptitude Tests

Early in the task analysis program while observing the Air Observer at work, it was noted that much of the job entails processing information from two or more sources at once, particularly in visual and auditory modes. This observation was substantiated in later stages of the analyses by the appearance of Dimensions I and II as defined in Tables 13 and 14. As Fournier (Note 18) states:

> For example, all crew members must monitor the intercommunication system while performing visual detection functions. Many of the work stations require the operator to manipulate equipment, monitor for targets, monitor for equipment malfunctions, and report status of detections while maintaining currency with the tactical scenario and crew communications (p. 1-2)

The fact that individuals respond differently when two or more physical or perceptual demands are made simultaneously than when either are presented singly has been noted for some time. For example, Chiles, Alluisi and Adams (1968), and Chiles and Jennings (Note 14) have suggested that individuals differ in their ability to "time-share" or "shift gears" from the requirements of one aspect to another. These authors have even implied that, in eliciting these differences, the nature of the task is not as important as the level of time sharing on the part of the operator.

Thus far, the effects on performance of having to "time-share" has been studied primarily in "dual-task" contexts in which information processing or action is required on stimuli presented simultaneously from both a primary and a secondary source. Performance measures taken under these conditions are compared to those taken when the stimuli from each source are presented separately. A drop in performance from the single-task situation to the dual-task situation is generally noted (see

Johnston, Greenberg, Fisher & Martin, 1970; Posner & Boies, 1971; Shulman & Greenberg, 1971; Smith, 1969; Taylor, Lindsay & Forbes, 1967). It has been suggested that the drop in performance from single-task presentation to dual-task presentation is inversely proportional to the "spare processing capacity" of the operator when he handles the primary task situation alone (Brown, 1962).

It is obvious from a number of the tasks loading at the low end of Dimension I that performance required simply to do a task is not necessarily what makes it complex for the Air Observer. The milieu, or what may be going on when the task is performed is also significant. For example, tasks 144 and 145 may not be complex to perform in and of themselves, but when they must be done under operational conditions, the situation can be complex. In this context, the individual must simultaneously process information from a variety of sources, as well as perform a number of other functions. Dimension II reflects even more directly a general requirement on the part of the Observer to handle dual or multi-source tasks. Tasks loadings at the high end of this dimension tend to be those in which an individual must accumulate, process and synthesize information (often received simultaneously) from several sources before making a decision.

On the basis of the evidence that the ability to simultaneously process information from more than one source was important to an Air Observer, a dual-task situation was created by presenting: (a) a primary task consisting of a number of slides each showing five pictures of aircraft in different orientations and attitudes along with readings on two aircraft instruments (artificial horizon and compass), and (b) a secondary task consisting of an auditorily presented series of random digits with a presentation rate of two seconds.

Forty-nine Observers (some of whom had participated in the task analysis study) were asked to select the aircraft picture corresponding to information presented on the instruments while repeating aloud, in sequence, one random digit after being given the next.

The psychometric qualities of the dual-task measures and their relation to on-job criteria in "concurrent validity comparisons" are presented in detail elsewhere (Fournier, Note 18). The following quotation from this source is provided as a succinct statement of some of these findings:

> Measures of the drop in performance (dual-task

> decrement) observed when the two tasks were combined compared to performance levels when done separately, showed that some Observers were able to perform in the dual-task situation better than others. The dual-task measures were not significantly related to operational experience...the Radar Simulator criterion...was significantly related to other criterion measures but did not appear to reflect a large dual-task component.
>
> Dual-task test measures were systematically (significantly) and positively related to job-related performance measures including supervisor rank ordering, peer ratings, final radar training grades, and three indices produced by combining the subjective and objective criteria measures (p. 11).

In adddition, though requiring confirmation by cross-validation, the evidence suggested that the dual-task measure offered a significant prediction increment when combined with selection procedures already in use.

The above example shows but one way the information from Tables 13 to 22 could be used as a basis for creating aptitude measures. Another strategy would be to examine tasks loading at either pole of specific content dimensions with a view to developing task replicas as measures of aptitude. In creating these instruments one would, of course, try to limit those tasks aspects that are dependent on specific previous learning. The practice of using work samples as predictors of later success in training and/or work situations is an established practice (see Cronbach, 1960). In fact, the only departure from tradition proposed here, is that the work sample would be selected on the basis of prior evidence that it contained a large component of a previously identified construct. More conventionally, job samples are generally introduced on a cut and try, intuitive basis. Then, if the resulting instrument predicts adequately, it survives. Often however, it is difficult to determine exactly what is being measured in these applications.

One notes from the interpretations given in Tables 13 to 22 that only Dimensions I, II, III, VII and perhaps VI seem to reflect aptitude-like qualities. Remaining dimensions appear to have more to do with task interrelationships and the milieu in which they are performed. Thus there is some evidence that the content dimensions are of different types. Participants in the study were asked to sort on the basis of how similar the tasks were to do. Therefore, given that attributes other than aptitude-related ones were relevant in making the sorting decisions, one would expect these to be reflected in the results.

Criterion dimensions identified and interpreted as in this study can be used to help decide what aptitude measures might be useful in a particular application. After this has been done, and the instruments prepared, one could use information arising from these dimensions to suggest the form and content of criterion data to use in validating the tests. There are a number of ways that the information from Tables 13-22 could assist in this process.

Criterion Data Collection Procedures

Evidence that reliable unidimensional scales can be generated from the dimensions produced in MDS studies has been provided by Schultz and Siegel (Note 16). These studies were conducted before the advent of recent MDS technology. As a consequence they were restricted to rather limited sets of stimuli. In spite of these limitations, these studies contain implications for significant extensions to the research methodology illustrated in the Air Observer research programme.

In one study for example, Schultz and Siegel (Note 10) used MDS to investigate content dimensions underlying successive interval judgements of 18 tasks associated with the trade of electronics technician in the U.S. Navy. The following four dimensions were produced: Electro-Comprehension, Equipment Operation and Inspection, Electro-Repair, and Electro-Safety. Taking dimension definitions derived from task loadings on each of the dimensions, the authors asked technicians to: (a) judge each task on the basis of its perceived relationship to each of the four dimensions; and (b) think of and evaluate other technicians on the task as viewed from the dimension definitions. From these judgements unidimensional scales were produced which met Thurstone and Guttman scaling requirements. For example, the indices of consistency I, which Green (1956) states should be .50 or higher before a set of items can be considered to scale in the Guttman sense were: Electro-Comprehension (.62), Equipment Operation and Inspection (.68), Electro-Repair (.74), and Electro-Safety (.77). Correlations between the direct task ratings on each of the defined dimensions and the task loadings on each dimension produced in the MDS analyses were Electro-Comprehension (.88), Operation and Inspection (.79), Electro-Repair (.67), and Electro-Safety (.50).

Generalizing from the Schultz and Siegel research program to that outlined in this paper has limitations because of the small number (18) of stimuli used in the former. The investigators were undoubtedly restricted by the number of variables their computer programmes could handle. The task analysis literature, however, (some of which was summarized earlier) suggests that this list of 18 tasks was either incomplete as a reflection of a skilled trade like Electronics Technician, or too general to serve as the basis of a task analysis in any realistic sense.

Following the lead of Schultz and Siegel (Note 16), one might use the definitions in Tables 13 to 22 as a basis for generating separate unidimensional scales. The correlations between task scale values on these scales and the task projections on the respective MDS dimensions would serve as indices of the adequacy of the scale development process. Large values would give one confidence in the unidimensional scales, attest to the validity of the MDS results, and support the individual interpretations ascribed. If some of the correlations were too small, one might try adjusting the scale definitions and then redoing the unidimensional scaling. Successive iterations with this strategy should sharpen the dimensional definitions. If the definitions of certain dimensions could not be brought into focus, this would serve as a cue that more study of the process or methodology used to generate them was required.

A number of conventional rating and ranking procedures (Torgerson, 1958) could be used to compare individual performance against the dimension definitions for purposes of collecting criteria data for test validation, and/or performance evaluation for promotions, job transfers, special assignments, and so on. Alternatively, one could investigate the feasibility of using task loadings on the content dimensions as the basis for "behaviourally based rating scales" analogous to those developed by Campbell, Dunnette, Arvey, and Hellervik (1973); Fogli, Hulin, and Blood (1971); Landy and Guion (1970), Smith and Kendall (1963); and Zedeck and Baker (1972). Since task scale values on each of the derived dimensions are already available, it would be possible to circumvent many of the scale development phases described by the above authors.

Suppose one had MDS content dimensions that were subsequently defined and redefined using Guttman procedures as described earlier in this section. Then to form a behaviourally based scale one would need only to select a number of tasks that were reasonably well distributed along the dimension. (Though perhaps not absolutely essential, it might be wise to select tasks loading only on the dimension of interest). Since the MDS/Guttman scaling procedures should have provided reasonably clear scale definitions, interpretation problems should be minimal if raters could be induced to concentrate (i.e., in comparing ratee performance to that required to do the task at a certain level of proficiency) only on the dimensional description of interest.

Conclusion

The research programme outlined in this paper was predicated on a contention that adequate technology for delineating performance/behavioural dimensions inherent in job tasks exists, and requires only to be organized and implemented in a systematic way. One set of procedures for doing this was presented by illustration in a task analysis of the Air Observer trade in the Canadian Forces. The results of these analyses: were reliable and internally consistent within relatively homogeneous groups of individuals; were readily and meaningfully generalizable across a variety of work situations (experience and responsibility levels); showed promise of being valid in terms of producing meaningful results showing significant relationship to external variables and being readily integrated into larger bodies of scientific knowledge; and had implications that could be extended in other studies.

There is no intent to imply that the methodology outlined represents a panacea for the "criterion problem" in its many facets. The goals and requirements of task analysis should change from application to application, necessitating corresponding adjustments in research methodology. The taxonomy and judgemental strategy (e.g., sorting or other procedures) for generating proximity indices, as well as the MDS models and other analytic techniques should be tailored to suit the specific application.

In Cronbach and Gleser's (1965) terms, the procedures outlined in this paper must be considered somewhat "narrow band" but hopefully of "high fidelity". Where data from "wider band" procedures were required (e.g., when investigating trade or occupational structures in large organizations), other kinds of methodologies are likely to be required. This qualification being granted, the evidence suggests that, taken together, the kinds of procedures used in the Air Observer task analysis represent a comprehensive, integrated research methodology not previously available. This methodology may not be universally appropriate, but when sensibly and appropriately applied, can produce reliable, internally consistent, and valid results of both theoretical and practical import.

Reference Notes

1. Thorndike, R. C. *Research problems and techniques* (Report No. 3). Washington, D.C.: U.S. Government Printing Office, Army Air Forces Aviation Psychology Programme Research Reports, 1947.

2. Crooks, D. S. *Criteria requirements for validation of the Classification Battery - Men* (Experimental) (Working paper). Toronto: Canadian Forces Personnel Applied Research Unit, 1974.

3. Kershner, A. M. *A report on job analysis* (ONR Report ACR-5). Washington, D.C.: Department of the Navy, 1955.

4. Ammerman, H. L. *A model of junior officer jobs for use in developing task inventories* (HumRRO Tech. Rep. 65-10). Alexandra, VA.: Human Resources Research Organization, November, 1965.

5. Dumas, N. S., & Muthard, J. E. Job analysis method for health-related professions: A pilot study of physical therapists. *Journal of Applied Psychology*, 1971, 55 (5), 458-465.

6. Finley, D. C., Obermayer, R. W., Bertone, C. M., Meister, D., & Muckler, F. A. *Human performance prediction in man-machine systems* (NASA CF-1614). Canoga Park, California: Bunder-Ramo Corporation, 1970.

7. Kruskal, J. B., Young, F. W., & Seery, J. B. *How to use KYST, a very flexible program to do multidimensional scaling and unfolding* (Bell Lab. Report). Murray Hill, [illegible]. J.: Bell Telephone Laboratories, 1972.

8. Rampton, G. M. *Task analysis in complex work environments* (Research Report 76-2). Kingston Ontario: Royal Military College of Canada, September, 1976.

9. Berliner, C., Angell, D., & Shearer, J. W. *Behaviours, measures, and instruments for performance evaluation in simulated environments*. Paper presented at the Symposium and Workshop on the Quantification of Human Performance, Albuquerque, New Mexico, 17-19 August, 1964.

10. Schultz, D. G., & Siegel, A. I. *Post-training performance criterion development and application: A multidimensional scaling analysis of the job performance of Naval aviation electronics technicians* Wayne, PA.: Applied Psychological Services, 1962.

11. Wheaton, G. R. *Development of a taxonomy of human performances: A review of classification systems relating to tasks and performance*

(Technical Report 1). Washington, C.C.: American Institutes for Research, 1968.

12. Siegel, A. I., & Schultz, D. G. *Post-training performance criterion development and application: A comparative multidimensional scaling analysis of the tasks performed by Naval aviation electronics technicians at two job levels*. Wayne, Pa.: Applied psychological Services, 1963.

13. Fournier, B. A. *Concurrent validation of a dual-task selection test for 081 Observers* (Report 74-1). Toronto: Canadian Forces Personnel Applied Research Unit, 1974.

14. Chiles, W. D., & Jennings, A. E. *Effects of alcohol on complex performance*. Washington, D.C.: FAA, Officer of Aviation Medicine Reports, 1969.

15. Schultz, D. G., & Siegel, A. I. *Post-training performance criterion development and application: The development of unidimensional scales for the dimensions derived from a multidimensional scale analysis of the job of the naval Aviation Electronics Technician*. Wayne, Pa.: Applied Psychological Services, 1964. (b)

References

Allusisi, E. A. *Principles of skill acquisition*. New York: Academic Press, 1969.

Brown, I. D. Measuring the "spare mental capacity" of car drivers by a subsidiary auditory task. *Ergonomics*, 1962, *5* (1), 247-250.

Brown, K. R. Job analysis by multidimensional scaling. *Journal of Applied Psychology*, 1967, *51* (6), 469-475.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. Development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, *57* (1), 15-22.

Chiles, W. D., Alluisi, E. A., & Adams, O. S. Work schedules and performance during confinement. *Human Factors*, 1968, *10*, 143-196.

Christensen J. M., & Mills, R. G. What does the operator do in complex systems? *Human Factors* 1967, *9* (4), 329-340.

Cronbach, L. J. *Essentials of psychological testing* (2nd ed.). New York: Harper & Row, 1960.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions* (2nd ed.). Urbana, Ill.: University of Illinois Press, 1965.

Dunnette, M. D. A note on the criterion. *Journal of Applied Psychology*, 1963, *47*, 251-254.

Fleishman, E. A. Development of a behaviour taxonomy for describing human tasks: A correlational experimental approach. *Journal of Applied Psychology*, 1967, *51* (1), 1-10. (a)

Fleishman, E. A. Individual differences and motor learning. In R. M. Gagne (Ed.), Learning and individual differences. Columbus, Ohio: Merrill Books, 1967. (b)

Fleishman, E. A. Performance assessment based on an empirically derived task taxonomy. Human Factors, 1967, 9 (4), 349-366 (c).

Fogli, L., Hulin, C. L., & Blood, M. R. Development of first level behavioral job criteria. Psychological Bulletin, 1971, 55, 3-8.

Gagne, R. M. Psychological principles in system development. New York: Holt, Rinehart, & Winston, 1963.

Garrett, H. E. Statistics in psychology and education. New York: David McKay, 1966.

Green, B. F. A method of scalogram analysis using summary statistics. Psychometrika, 1956, 21, 79-88.

Grodsky, M. A. The use of full scale mission simulation for the assessment of complex operator performance. Human Factors, 1967, 9 (4), 341-348.

Inn, A., Hulin, C. L., & Tucker, L. Three sources of criterion variance: Static dimensionality, dynamic dimensionality, and individual dimensionality. Organizational Behaviour and Human Performance, 1972, 8, 58-83.

Johnston, W. A., Greenberg, S. N., Fisher, R. D., & Martin, D. W. Divided attention: A vehicle for monitoring memory process. Journal of Experimental Psychology, 1970, 83 (1), 164-171.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 1-28. (a)

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964, 29 115-130. (b)

Landis, D., Silver, C. A., Jones, J. M., & Messick, S. Level of proficiency and multidimensional viewpoints about problem similarity. *Journal of Applied Psychology*, 1967, *51* (3), 216-222.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, *5*, 93-103

Miller, J. E., Shephard, R. N., & Chang, J. J. An analytical approach to the interpretation of multidimensional scaling solutions. *American Psychologist*, 1964, *19*, 579-580 (Abstract).

Miller, R. B. Task description and analysis. In R. M. Gagne (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart & Winston, 1963.

Posner, M. I., & Boies, S. J. Components of attention. *Psychological Bulletin*, 1971, *78* (5), 391-408.

Prien, D. P., & Ronan, W. W. Job analysis: A review of research findings. *Personnel Psychology*, 1971, *24* 371-396.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, *27* (3), 219-246.

Shulman, H. G., & Greenberg, S. N. Perceptual deficit due to division of attention between memory and perception. *Journal of Experimental Psychology*, 1971, *88* (2), 171-176.

Smith, M. C. Effect of varying channel capacity on stimulus detection and discrimination. *Journal of Experimental Psychology*, 1969, *82* (3), 520-526.

Smith, P., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, *47*, 149-155.

Smith, R. J., & Siegel, A. I. A multidimensional scaling analysis of the job of Civil Defence Director. Journal of Applied Psychology, 1967, 51 (6), 476-480.

Spence, I. A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. Psychometrika, 1972, 37 (4), 461-486.

Taylor, M. M., Lindsay, P. H., & Forbes, S. M. Quantification of shared capacity processing in auditory and visual discrimination. Acta Psychologica, 1967, 27, 223-229.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Vernon, P. E. Ability factors and environmental influences. American Psychologist, 1965, 20, 723-733.

Young, F. W., & Torgerson, W. S. TORSCA, a FORTRAN IV program for Shepard-Kruskal multidimensional scaling analysis. Behavioral Science, 1967, 12, 498.

Zedeck, S., & Baker, H. Evaluation of behavioral expectation scales. Paper presented at the meeting of the Midwestern Psychological Association, Detroit, May 1971.

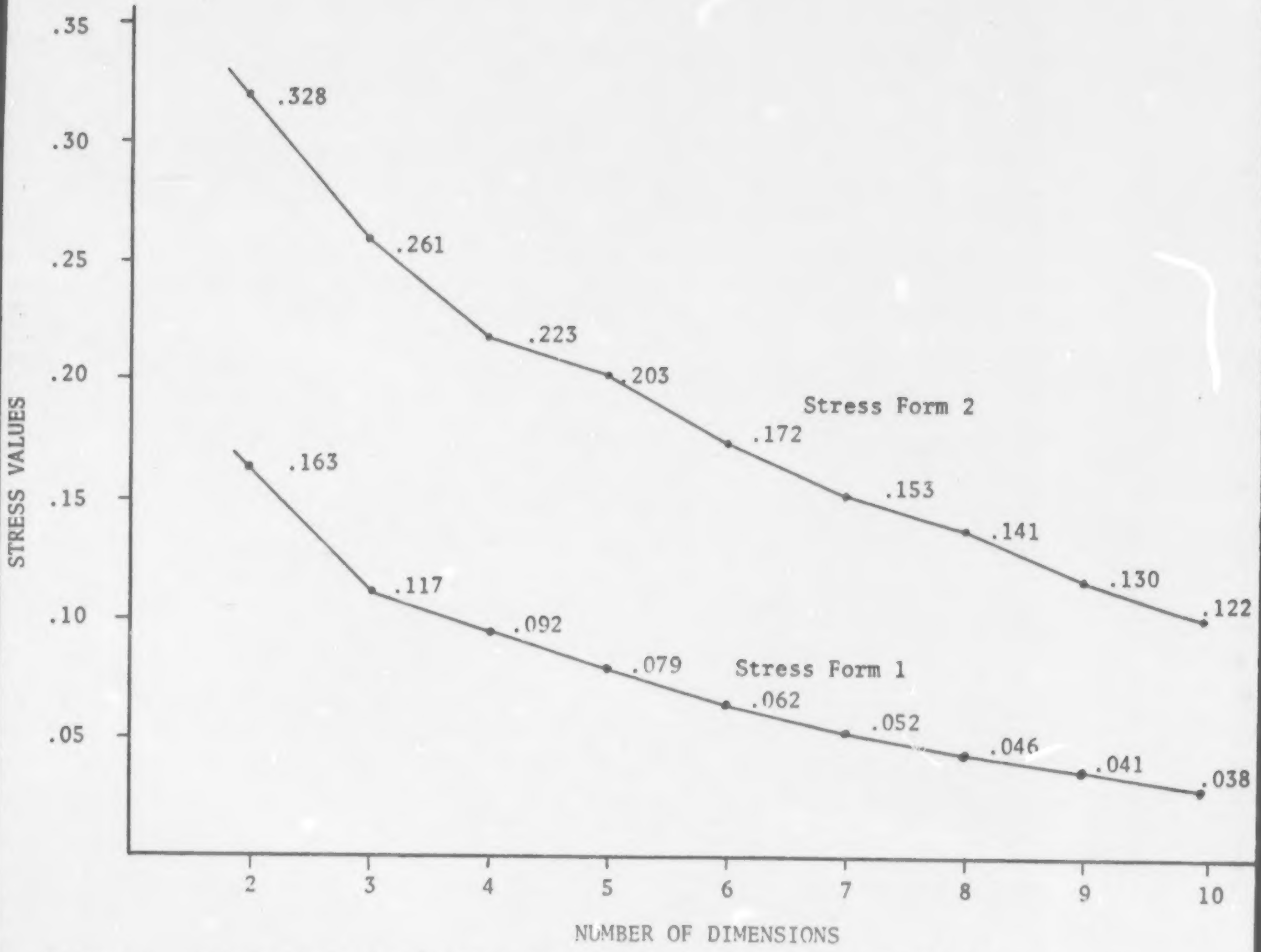
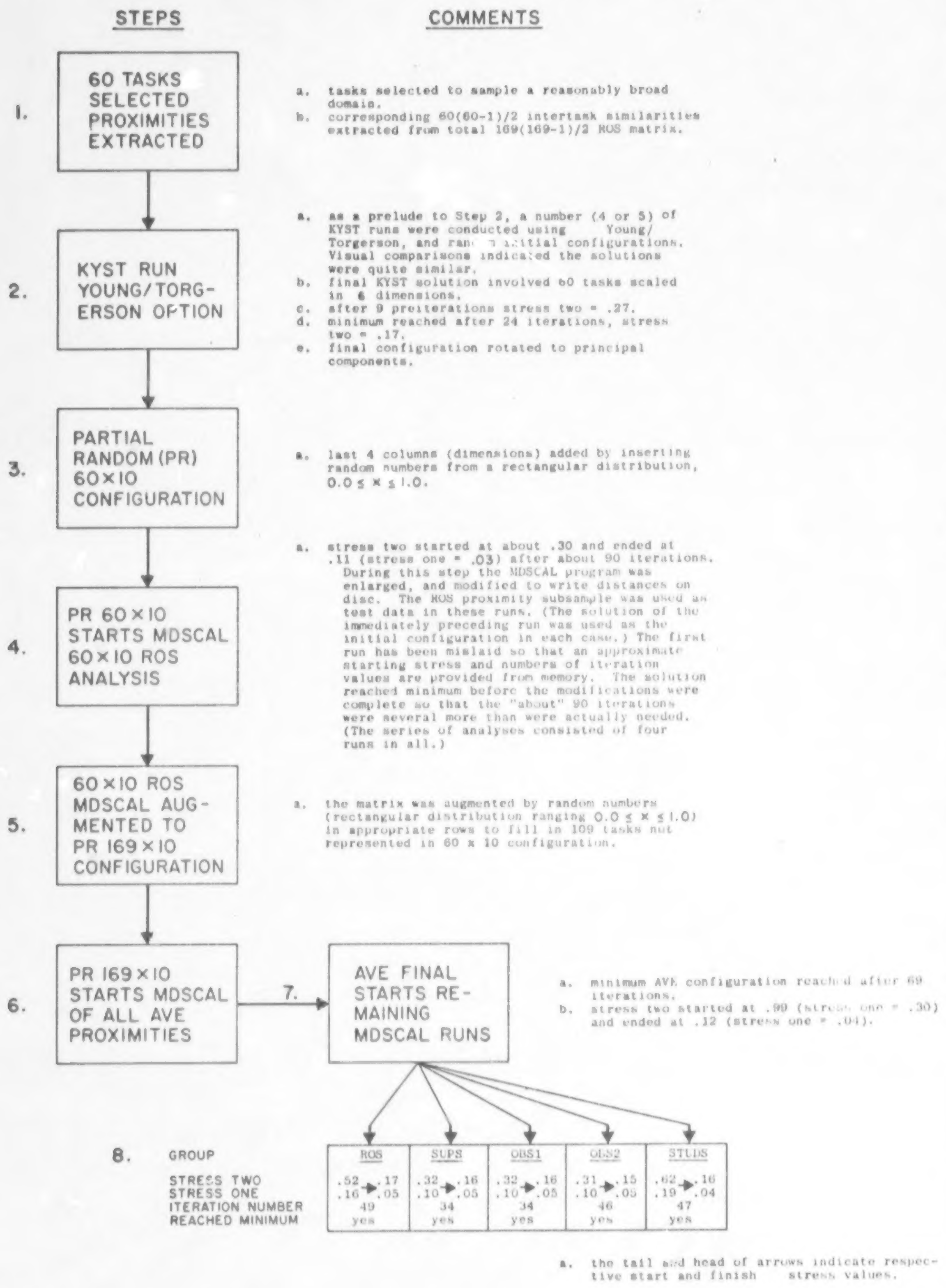


Figure 1. Stress plots for MDSCAL analyses of AVE proximity data in ten to two dimensions.


STEPS
COMMENTS
1.
60 TASKS SELECTED PROXIMITIES EXTRACTED
2.
KYST RUN YOUNG/TORGERSON OPTION
3.
PARTIAL RANDOM (PR) 60×10 CONFIGURATION
4.
PR 60×10 STARTS MDSCAL 60×10 ROS ANALYSIS
5.
60×10 ROS MDSCAL AUGMENTED TO PR 169×10 CONFIGURATION
6.
PR 169×10 STARTS MDSCAL OF ALL AVE PROXIMITIES
7.
AVE FINAL STARTS REMAINING MDSCAL RUNS
8.
GROUP
STRESS TWO
STRESS ONE
ITERATION NUMBER
REACHED MINIMUM
ROS
.52 .17
.16 .05
49
yes
SUPS
.32 .16
.10 .05
34
yes
OBS1
.32 .16
.10 .05
34
yes
OBS2
.31 .15
.10 .05
46
yes
STUDS
.62 .16
.19 .04
47
yes

1.

a. tasks selected to sample a reasonably broad domain.
b. corresponding 60(60-1)/2 intertask similarities extracted from total 169(169-1)/2 ROS matrix.

2.

a. as a prelude to Step 2, a number (4 or 5) of KYST runs were conducted using Young/Torgerson, and random initial configurations. Visual comparisons indicated the solutions were quite similar.
b. final KYST solution involved 60 tasks scaled in 6 dimensions.
c. after 9 preiterations stress two = .27.
d. minimum reached after 24 iterations, stress two = .17.
e. final configuration rotated to principal components.

3.

a. last 4 columns (dimensions) added by inserting random numbers from a rectangular distribution, $0.0 \leq x \leq 1.0$.

4.

a. stress two started at about .30 and ended at .11 (stress one = .03) after about 90 iterations. During this step the MDSCAL program was enlarged, and modified to write distances on disc. The ROS proximity subsample was used as test data in these runs. (The solution of the immediately preceding run was used as the initial configuration in each case.) The first run has been mislaid so that an approximate starting stress and numbers of iteration values are provided from memory. The solution reached minimum before the modifications were complete so that the "about" 90 iterations were several more than were actually needed. (The series of analyses consisted of four runs in all.)

5.

a. the matrix was augmented by random numbers (rectangular distribution ranging $0.0 \leq x \leq 1.0$) in appropriate rows to fill in 109 tasks not represented in 60 x 10 configuration.

7.

a. minimum AVE configuration reached after 69 iterations.
b. stress two started at .99 (stress one = .30) and ended at .12 (stress one = .04).

8.

a. the tail and head of arrows indicate respective start and finish stress values.

Table 1. Schematic summary of MDSCAL analyses.

Table 2

Multiple Linear Regression of Each ROS Dimension on SUPS Dimensions

Multiple Correlations

ROS Dimensions	I	II	III	IV	V	VI	VII	VIII	IX	X
Multiple Correlations	.93	.88	.88	.68	.87	.75	.77	.87	.75	.67

Matrix of Direction Cosines Showing Correspondence Between Fitted ROS Vectors and SUPS Dimensions

SUPS Dimensions	I	II	III	IV	V	VI	VII	VIII	IX	X
I	*.81*	-.06	.11	-.03	.12	.07	.04	.06	-.02	-.05
II	-.31	*.94*	.05	.04	.13	-.12	-.21	-.23	.02	-.13
III	-.13	-.18	*.90*	.10	.26	-.02	-.14	-.14	.07	.24
IV	-.36	-.12	.01	*.90*	.20	-.01	.02	-.27	.05	-.04
V	-.07	.04	.18	.15	*.85*	.14	.02	-.09	.05	.03
VI	.09	-.13	-.00	.12	.04	*.93*	.11	-.05	-.06	.03
VII	-.13	-.16	-.02	-.01	.05	-.10	*.94*	-.12	-.00	.04
VIII	.25	-.08	-.35	.12	-.28	-.18	.02	*.90*	-.02	.10
IX	-.08	.11	.07	.28	-.12	-.20	-.17	-.09	*.98*	-.06
X	-.05	-.01	.01	-.22	.22	-.06	-.03	.03	-.14	*.95*

Matrix of Cosines Showing Relationships Between ROS Vectors After Being Fitted into SUPS Configuration

ROS Dimensions										
I	-									
II	-.30	-								
III	-.16	-.08	-							
IV	-.35	-.06	.07	-						
V	-.18	.08	.52	.22	-					
VI	.17	-.22	.06	.06	.19	-				
VII	-.02	-.42	.08	.01	.09	.07	-			
VIII	.49	-.23	-.46	.21	-.44	-.16	-.06	-		
IX	-.14	.12	.13	.36	-.07	-.23	-.17	-.14	-	
X	-.04	-.20	.28	-.22	.25	-.03	.09	.11	-.19	-

Table 3

Multiple Linear Regression of Each ROS Dimension on OBS1 Dimensions

ROS Dimensions

Multiple Correlations

I	II	III	IV	V	VI	VII	VIII	IX	X
.93	.87	.88	.76	.86	.77	.81	.83	.66	.74

Matrix of Direction Cosines Showing Correspondence Between Fitted ROS Vectors and OBS1 Dimensions

OBS1 Dimensions	I	II	III	IV	V	VI	VII	VIII	IX	X
I	*.87*	-.02	.06	-.08	.02	.11	-.03	.07	-.02	-.05
II	-.14	*.94*	-.14	.05	.11	-.05	-.13	-.15	-.03	-.19
III	-.07	-.14	*.91*	-.03	.15	-.04	.06	-.05	.11	.04
IV	-.29	-.05	-.11	*.90*	.19	.04	.08	-.26	.06	-.08
V	-.18	.04	.19	.14	*.92*	.11	.00	-.17	.04	.15
VI	.02	-.20	-.01	.22	.14	*.91*	.07	-.12	-.07	-.02
VII	-.08	-.13	-.04	.05	.00	-.06	*.98*	-.03	.10	.01
VIII	.29	-.09	-.27	.10	-.08	-.33	.00	*.93*	-.06	.01
IX	.02	-.07	.14	.27	.04	-.19	.03	.00	*.90*	-.04
X	-.12	-.16	.02	-.13	.24	-.04	.07	.02	-.21	*.96*

Matrix of Cosines Showing Relationships Between ROS Vectors After Being Fitted into OBS1 Configuration

ROS Dimensions										
I	-									
II	-.14	-								
III	-.07	-.22	-							
IV	-.32	-.04	-.11	-						
V	-.28	.05	.31	.30	-					
VI	.00	-.17	.05	.16	.24	-				
VII	-.12	-.28	.02	.13	.05	.00	-			
VIII	.47	-.19	-.27	-.21	-.32	-.42	-.04	-		
IX	-.03	-.07	.25	.33	.05	-.22	.11	-.07	-	
X	-.14	-.32	.10	-.20	.32	-.04	.10	.05	-.23	-

Table 4

Multiple Linear Regression of Each ROS Dimension on OBS2 Dimensions

ROS Dimensions

Multiple Correlations

I	II	III	IV	V	VI	VII	VIII	IX	X
.95	.86	.89	.74	.86	.75	.80	.86	.76	.67

Matrix of Direction Cosines Showing Correspondence Between Fitted ROS Vectors and OBS2 Dimensions

OBS1 Dimensions	I	II	III	IV	V	VI	VII	VIII	IX	X
I	*.90*	-.07	-.03	.01	.02	.10	.01	.10	-.04	.00
II	-.20	*.92*	-.00	.06	.12	.03	-.04	-.05	.16	-.06
III	-.08	-.12	*.99*	-.04	.11	.04	.02	-.00	.06	.24
IV	-.17	-.04	-.02	*.95*	.14	.09	.03	-.09	.07	-.25
V	-.01	.18	.04	.00	*.94*	.17	-.08	-.07	.08	-.12
VI	.15	-.00	.11	-.07	.09	*.91*	-.00	-.16	.02	-.11
VII	-.10	-.30	-.07	.18	.04	.01	*.99*	.05	.13	.27
VIII	.28	-.00	-.05	-.09	-.16	-.21	-.02	*.96*	.12	-.02
IX	.07	.10	.03	.16	-.13	-.27	.04	.13	*.95*	.08
X	.08	-.02	-.00	-.18	.17	-.10	.04	-.07	-.15	*.88*

Matrix of Cosines Showing Relationship Between ROS Vectors After Being Fitted into OBS2 Configuration

ROS Vectors										
I	-									
II	-.20	-								
III	-.09	-.09	-							
IV	-.22	-.02	-.07	-						
V	-.07	.23	.16	.10	-					
VI	.12	.01	.15	.02	.31	-				
VII	-.07	-.35	-.05	.20	-.04	-.03	-			
VIII	.35	-.07	-.07	-.11	-.28	-.39	.09	-		
IX	-.01	.22	.08	.26	-.05	-.23	.16	.24	-	
X	.07	-.19	.20	-.34	.01	-.24	.32	-.01	-.05	-

Table 5

Multiple Linear Regression of Each ROS Dimension on STUDS Dimensions

ROS Dimensions

	I	II	III	IV	V	VI	VII	VIII	IX	X
Multiple Correlations	.91	.78	.79	.71	.77	.53	.76	.82	.60	.65

Matrix of Direction Cosines Showing Correspondence Between Fitted ROS Vectors and STUDS Dimensions

STUDS Dimension	I	II	III	IV	V	VI	VII	VIII	IX	X
I	*.87*	-.04	-.05	.03	-.06	.14	-.06	.32	.09	-.14
II	-.27	*.84*	-.00	-.01	-.03	-.22	-.41	-.24	.05	-.15
III	-.23	-.15	*.92*	.02	.33	.11	.07	-.22	.27	.19
IV	-.17	-.18	.04	*.89*	.12	-.04	.16	-.29	-.08	.08
V	-.23	.02	.32	-.12	*.88*	-.08	.01	-.25	.10	.17
VI	.03	-.16	.17	.14	.22	*.86*	.13	-.39	-.12	.09
VII	.03	-.29	-.15	18	.05	-.09	*.78*	.11	-.00	.02
VIII	.10	-.22	-.03	.00	.06	-.35	.21	*.68*	-.16	.19
IX	.10	.05	.02	.37	-.17	-.13	-.28	-.06	*.90*	-.23
X	.10	-.27	-.00	.02	.09	-.17	.23	.13	-.23	*.89*

Matrix of Cosines Showing Relationships Between ROS Vectors After Being Fitted into STUDS Configuration

ROS Vectors										
I	-									
II	-.26	-								
III	-.34	-.13	-							
IV	-.05	-.24	.02	-						
V	-.34	-.18	.62	-.01	-					
VI	.14	-.20	.20	.03	.12	-				
VII	.06	-.75	-.02	.20	.22	.05	-			
VIII	.57	-.29	-.37	-.28	-.36	-.49	.23	-		
IX	.03	.17	.27	.23	-.04	-.10	-.37	-.19	-	
X	-.09	-.47	.25	-.00	.38	-.10	.43	.11	-.40	-

Table 6

Multiple Linear Regression of Each OBS2 Dimension on OBS1 Dimensions

OBS2 Dimensions

Multiple Correlations

I	II	III	IV	V	VI	VII	VIII	IX	X
.98	.92	.93	.91	.92	.88	.89	.90	.88	.84

Matrix of Direction Cosines Showing Correspondence Between Fitted OBS2 Vectors and OBS1 Dimensions

OBS1 Dimensions	I	II	III	IV	V	VI	VII	VIII	IX	X
I	*.96*	-.01	-.01	.01	.01	.01	-.04	.03	.01	-.06
II	-.03	*.95*	-.19	-.13	-.11	.07	-.09	-.01	-.20	-.64
III	-.04	-.14	*.95*	.03	.14	-.03	-.02	-.10	.10	.05
IV	-.12	-.08	-.02	*.97*	.02	.08	.10	-.17	.02	.07
V	-.08	-.12	.16	-.03	*.97*	-.16	.12	-.10	-.00	.05
VI	.09	-.05	-.06	.04	-.01	*.95*	.07	-.09	.00	-.12
VII	-.16	-.04	-.01	.07	.02	.12	*.94*	-.05	-.04	.12
VIII	.12	-.03	-.15	-.15	-.04	-.14	-.01	*.96*	-.12	-.02
IX	-.00	-.23	.07	.12	.09	-.03	-.09	-.10	*.95*	-.14
X	-.13	-.04	.08	-.02	.13	-.14	.25	.09	-.14	*.97*

Matrix of Cosines Showing Relationship Between OBS2 Vectors After Being Fitted into OBS1 Configuration

OBS2 Dimensions										
I	-									
II	-.01	-								
III	-.08	-.34	-							
IV	-.13	-.23	.06	-						
V	-.10	-.27	.33	.01	-					
VI	.08	.05	-.11	.13	-.19	-				
VII	-.23	-.14	.02	.16	.17	.13	-			
VIII	.16	.03	-.25	-.33	-.15	-.23	-.05	-		
IX	.01	-.41	.21	.18	.10	-.01	-.14	-.24	-	
X	-.23	-.09	.15	.06	.18	-.23	.39	.06	-.25	-

Table 7

Multiple Linear Regression of Each Rated Property on AVE Dimensions

Properties

	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
Multiple Correlations	.33	.26	.27	.34	.49	.40	.33

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and AVE Dimensions

Configuration Dimensions	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
I	-.11	-.01	.14	-.19	.27	-.20	-.03
II	.25	.43	-.65	-.19	-.09	-.22	.37
III	.11	.05	.12	.20	.02	.19	.17
IV	.02	.09	-.14	.05	-.17	.02	.11
V	-.00	.22	-.40	-.17	-.21	-.41	.15
VI	.59	.49	-.49	.38	-.50	.45	.67
VII	-.12	-.36	.09	-.12	.26	-.09	-.30
VIII	.38	.25	-.15	.15	.15	.34	.25
IX	.51	.56	-.10	.45	.63	.38	.43
X	-.38	-.13	-.29	-.69	.33	-.49	-.14

Cosines of Angles Between Property Vectors in Configuration

Properties							
1. Concentration	-						
2. Difficulty	.88	-					
3. Manual Skill	-.47	-.70	-				
4. Importance	.78	.51	.11	-			
5. Cooperation	-.12	-.10	.29	-.15	-		
6. Speed	.77	.44	.12	.93	-.07	-	
7. Mental Effort	.92	.96	-.69	.57	-.25	.55	-

Table 8

Multiple Linear Regression of Each Rated Property on ROS Dimensions

	Properties						
	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
Multiple Correlations	.30	.27	.25	.29	.48	.36	.30

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and ROS Dimensions

Configuration Dimensions	1	2	3	4	5	6	7
I	-.10	-.11	.25	-.21	.36	-.11	-.11
II	.63	.67	-.42	.10	.05	.04	.72
III	.10	.12	.24	.15	.05	.28	.13
IV	.18	.16	.41	.62	.27	.44	.24
V	-.30	-.22	-.59	-.45	-.14	-.66	-.20
VI	.40	.40	-.36	.17	-.78	.44	.44
VII	.14	.02	.11	-.06	.21	-.11	.03
VIII	.27	.29	-.12	.06	.00	.23	.23
IX	.32	.42	-.18	.17	.06	-.02	.31
X	-.32	-.17	-.11	-.51	.34	-.16	-.07

Properties	Cosines of Angles Between Property Vectors in Configuration						
1. Concentration	-						
2. Difficulty	.97	-					
3. Manual Skill	-.20	-.32	-				
4. Importance	.65	.55	.41	-			
5. Cooperation	-.28	-.27	.52	.16	-		
6. Speed	.61	.55	.41	.82	-.24	-	
7. Mental Effort	.95	.98	-.31	.53	-.24	.57	-

Table 9

Multiple Linear Regression of Each Rated Property on SUPS Dimensions

Properties

	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
Multiple Correlations	.53	.30	.28	.28	.46	.32	.35

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and SUPS Dimensions

Configuration Dimensions	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
I	.01	.15	.22	-.00	.41	-.15	.11
II	.24	.28	-.68	-.33	-.00	-.24	.30
III	.00	.06	.09	.14	.02	.05	.12
IV	-.11	-.13	-.33	-.24	-.22	.01	-.03
V	.21	.38	-.26	-.00	-.13	-.32	.31
VI	.76	.51	-.42	.44	-.60	.70	.69
VII	-.38	-.61	-.08	-.48	.30	-.33	-.51
VIII	.21	.09	-.08	-.20	.09	.21	.05
IX	.26	.32	-.21	.23	.41	.22	.21
X	-.22	.01	-.27	-.54	.39	-.34	.04

Cosines of Angles Between Property Vectors in Configuration

Properties							
1. Concentration	-						
2. Difficulty	.88	-					
3. Manual Skill	-.47	-.45	-				
4. Importance	.61	.52	.28	-			
5. Cooperation	-.53	-.30	.23	-.48	-		
6. Speed	.71	.44	-.02	.75	-.56	-	
7. Mental Effort	.92	.96	-.54	.50	-.45	.52	-

Table 10

Multiple Linear Regression of Each Rated Property on OBS1 Dimensions

	Properties						
	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
Multiple Correlations	.31	.23	.27	.39	.49	.36	.31

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and OBS1 Dimensions

Configuration Dimensions	1	2	3	4	5	6	7
I	-.14	-.01	.13	-.19	.25	-.19	-.05
II	.41	.69	-.57	-.06	-.19	-.20	.52
III	.10	-.05	.15	.22	.09	.20	.16
IV	.26	.42	-.01	.24	-.08	.16	.38
V	-.06	.18	-.45	-.08	-.24	-.43	.14
VI	.46	.21	-.53	.37	-.50	.44	.53
VII	-.07	-.34	.11	-.11	.26	-.12	-.30
VIII	.31	.01	-.15	.20	.31	.24	.15
IX	.42	.38	-.03	.43	.56	.37	.31
X	-.49	-.12	-.35	-.69	.33	-.52	-.22

Properties	Cosines of Angles Between Property Vectors in Configuration						
1. Concentration	-						
2. Difficulty	.72	-					
3. Manual Skill	-.35	-.60	-				
4. Importance	.85	.40	.07	-			
5. Cooperation	-.19	-.23	.37	-.16	-		
6. Speed	.73	.19	.20	.91	-.05	-	
7. Mental Effort	.88	.89	-.61	.64	-.35	.48	-

Table 11

Multiple Linear Regression of Each Rated Property on OBS2 Dimensions

Properties

1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.

Multiple Correlations

1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
.34	.27	.26	.32	.48	.41	.33

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and OBS2 Dimensions

Configuration Dimensions	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
I	-.07	.07	.26	-.19	.27	-.09	-.01
II	.07	.34	-.69	-.37	-.13	-.35	.25
III	.24	.26	.00	.22	-.01	.33	.34
IV	-.13	-.11	-.27	-.06	-.24	-.15	-.06
V	-.20	.03	-.40	-.41	-.23	-.59	-.08
VI	.74	.64	-.39	.26	-.43	.39	.75
VII	-.31	-.37	.07	-.16	.20	-.16	-.31
VIII	.24	.11	-.20	-.13	.11	.11	.15
IX	.37	.48	.12	.20	.64	.18	.35
X	-.16	-.03	.08	-.67	.39	-.42	-.07

Properties	Cosines of Angles Between Property Vectors in Configuration						
1. Concentration	-						
2. Difficulty	.90	-					
3. Manual Skill	-.28	-.45	-				
4. Importance	.53	.24	.27	-			
5. Cooperation	-.13	-.05	.59	-.19	-		
6. Speed	.70	.37	.30	.90	-.04	-	
7. Mental Effort	.96	.96	-.43	.37	-.17	.53	-

Table 12

Multiple Linear Regression of Each Rated Property on STUDS Dimensions

Properties

1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.

Multiple Correlations

1	2	3	4	5	6	7
.31	.22	.22	.29	.43	.32	.31

Matrix of Direction Cosines Showing Correspondence Between Seven Fitted Property Vectors and STUDS Dimensions

Configuration Dimensions	1 Conc.	2 Diff.	3 MS.	4 Imp.	5 Coop.	6 Speed	7 ME.
I	.07	.18	.10	-.12	.32	-.07	.19
II	.59	.56	-.67	.26	-.03	.15	.58
III	.23	.38	-.38	.27	-.06	.28	.43
IV	.00	.07	-.23	.09	.09	.11	.01
V	-.18	-.12	-.17	.00	-.15	-.60	-.01
VI	.19	.11	-.13	.44	-.26	.37	.21
VII	-.11	-.38	.36	-.15	.36	-.07	-.28
VIII	-.03	-.15	.32	.00	.27	.15	-.24
IX	.68	.55	.25	.67	.35	.50	.49
X	-.23	-.08	-.07	-.41	.68	-.34	-.10

Cosines of Angles Between Property Vectors in Configuration

Properties	1	2	3	4	5	6	7
1. Concentration	-						
2. Difficulty	.91	-					
3. Manual Skill	-.34	-.55	-				
4. Importance	.86	.74	-.22	-			
5. Cooperation	.00	-.03	.37	-.28	-		
6. Speed	.74	.60	-.01	.78	-.08	-	
7. Mental Effort	.90	.97	-.61	.75	-.10	.54	-

Table 13

AVE Dimension I: Complexity of Task Context

Task Number	Task Statement	Scale Value
	Context Straightforward (task itself not necessarily easy)	
105	PREPARING MARKER MARINE FOR LAUNCH (ASW)	1.05
108	DROPPING PARACHUTE FLARE (ASW)	1.01
100	KEEPING CHUTES LOADED IAW NAVS INSTRUCTIONS (ASW)	1.01
098	FIRING SONO CHUTES LOCALLY (ASW)	1.00
106	LAUNCHING MARKER MARINE (ASW)	.99
103	FIRING LIBRASCOPE MANUALLY (ASW)	.99
117	SETTING UP MAI BAGS	.98
095	CHECKING SONO CHUTES ON PFI (ASW)	.98
110	CHECKING HULCHER CAMERA SERVICEABILITY (FRAME COUNTER & MOTOR VIBRATION)	.97
096	SETTING SWITCHES ON SONOS FOR PROPER DEPTH/LIFE (ASW)	.96

Comments

Tasks loading on the positive end of this dimension tend to involve fairly gross physical actions. Many of these are relatively straightforward, and little judgement is required when to initiate action, since the stimulus to initiate action generally originates from outside the individual, often from another member of the crew.

The situation or milieu in which the action can be performed is obviously important. For example, tasks 144 and 145 may not be complex to perform in and of themselves, but when they have to be done under the pressure of an operational mission, the larger work context of which they are a part can take on immense complexity.

	Context Complex (task itself not necessarily difficult)	
021	CHANGING RANGE SCALE DURING HOMING	-.79
152	CHECKING SERVICEABILITY OF ARC505 WITH EXTERNAL AGENCY ON VOICE (COMM)	-.79
058	SELECTING & ADJUSTING CONTROLS FOR MAD OPERATION (DETECTION)	-.81
149	CALIBRATING MODEM (COMM)	-.82
150	CHECKING SERVICEABILITY OF ORESTES CONTROL BOX (COMM)	-.84
145	CHANGING PAPER, RIBBONS & TAPE IN TELEPRINTERS (COMM)	-.85
151	CHECKING SERVICEABILITY OF JASON CONTROL BOX (COMM)	-.86
144	CHECKING PAPER, RIBBONS & TAPE IN TELEPRINTERS (COMM)	-.89
017	ADVISING PILOT OF HEADING TO PERFORM HOMING (ASV)	-.91

Table 14

AVE Dimension II: Multivariate Nature of Task and Degree of Interpretation/Decision Making When Performing Task

Task Number	Task Statement	Scale Value
	Important Interpretation/Decision Making Component, Typically with Many Facets	
072	OBTAINING TURN COUNT (AURAL LISTENING-DETECTION)	.57
071	CATEGORIZING TARGET BY CLASS, DOPPLER, DISTANCE (AURAL LISTENING-DETECT)	.56
025	MAINTAINING VARIABLE RANGE MARKER ON TARGET (ASV)	.54
027	CALLING ACCURATE ON-TOP (ASV)	.53
029	INSPECTING MAP FOR IDENTIFIABLE LANDMARKS (ASV)	.51
028	ADJUSTING SCOPE PRESENTATION FOR BEST MAP READING (ASV)	.51
070	DISCRIMINATING TARGET FROM BACKGROUND (AURAL LISTENING-DETECTION)	.48
010	CHECKING QUALITY OF SCOPE PRESENTATION ON GROUND VIDEO CHECK (ASV)	.47
033	INFORMING NAVIGATOR OF LANDMARK & ITS RANGE & BEARING (ASV)	.47
008	SETTING UP GIVEN SECTOR FOR TRANSMITTER CHECK (ASV)	.46

Comments

Associated with the high end of this dimension are tasks with much interpretation/decision making, generally involving many different variables.

The AVE configuration vectors relating most strongly to difficulty, manual skill, and mental effort have respective direction cosines of .43, -.65, and .37 with this dimension. This profile is consistent with the interpretation ascribed this scale.

Task Number	Task Statement	Scale Value
	Straightforward Tasks with Low Level of Interpretation/Decision Making	
107	PREPARING PARACHUTE FLARE FOR DROP (ASW)	-.47
154	MAKING A MESSAGE TAPE (COMM)	-.48
120	PFI & PRESETTING OF ECM COMPONENTS	-.50
109	RETURNING PARACHUTE FLARE TO STORAGE (ASW)	-.54
108	DROPPING PARACHUTE FLARE (ASW)	-.54
002	CHECKING METER VOLTAGES ON TURN-ON (ASV)	-.55
093	FIRING PETRO LOCALLY (ASW)	-.55
119	ACTING AS A MAI DROP MASTER	-.56
118	DROPPING MAI BAGS IN SEQUENCE	-.64
117	SETTING UP MAI BAGS	-.67

201

Table 15

AVE Dimension III: Fineness/Grossness of Task Activity

Task Number	Task Statement	Scale Value
	Manipulating, Finetuning, Checking Kinds of Behaviours	
089	IDENTIFYING & RECTIFYING AP102 FAULTS LISTED IN CHECKLIST (DETECTION)	.93
061	IDENTIFYING & RECTIFYING MAD FAULTS LISTED IN CHECKLIST (DETECTION)	.93
078	IDENTIFYING & CALLING ECHOES DURING JULIE OPERATIONS (DETECTION)	.87
077	SELECTING PROPER BUOYS DURING JULIE OPERATIONS (DETECTION)	.86
020	CALLING RANGES DURING HOMING (ASV)	.80
088	SETTING UP AR102 REMOTE CONTROL TO RECORD A FACILITY (DETECTION)	.79
073	INSPECTING JULIE RX & AJH501 RECORDERS DURING PFI (DETECTION)	.76
056	CHECKING MAD INTERNAL NOISE (DETECTION)	.74
069	LISTENING TO WATER AND TARGET AUDIO (AURAL LISTENING-DETECTION)	.73
040	CHANGING ASV RECEIVER CRYSTALS (ASV)	.73

Comments

Tasks relating to checking, fine manipulation, tuning, identifying, and so on tend to load highly on this dimension, whereas tasks requiring gross response behaviours predominate on the lower end. While this dimension in the STUDS configuration share direction cosines of .38, -.38 and .43 with Difficulty, Manual Skill, and Mental Effort, respectively, relationships with any of the seven properties in configurations of the remaining groups were generally quite small.

	Gross Behaviours, Heavy Lifting, Carrying, General Dogwork	
136	IDENTIFYING BASIC RADAR TYPE FROM AURAL PRF (ECM)	-.44
114	TAKING PICTURES IN NOSE WITH HULCHER CAMERA	-.44
141	ADJUSTING AE GAIN AND/OR ATTENUATION DURING HOMING (ECM)	-.45
115	KEEPING ACCURATE HULCHER CAMERA LOG (ASW)	-.45
164	PRESSING SYNCH BUTTON ON TIME CHECK CUE ON LF (COMM)	-.46
169	IDENTIFYING AND RECTIFYING COMM EQUIPMENT FAULTS LISTED IN CHECKLIST	-.46
093	FIRING RETRO LOCALLY (ASW)	-.48
130	ADJUSTING SCOPE FOR BEST D/F SIGNAL (ECM)	-.50
094	UNLOADING & TURNING OFF RETRO (ASW)	-.52
132	READING ANALYZER FOR PW & PRF (ECM)	-.54

Table 16

AVE Dimension IV: Criticality of Task Activity to Mission Success

Task Number	Task Statement	Scale Value
	Important to Successful Completion of Mission That Task Done Correctly	
086	LOADING TAPES IN AR102 TAPE RECORDERS (DETECTION)	.49
163	KEYING JASON (COMM)	.48
077	SELECTING PROPER BUOYS DURING JULIE OPERATIONS (DETECTION)	.45
149	CALIBRATING MODEM (COMM)	.44
087	CHECKING AR102 RECORDER METER FOR RECORDING ON BOTH CHANNELS (DETECTION)	.43
078	IDENTIFYING & CALLING ECHOES DURING JULIE OPERATIONS (DETECTION)	.42
067	PLACING RULER ON CHART FOR TARGET RANGE CHECK (SSQ47-DETECTION)	.41
150	CHECKING SERVICEABILITY OF ORESTES CONTROL BOX (COMM)	.41
088	SETTING UP AR102 REMOTE CONTROL TO RECORD A FACILITY (DETECTION)	.40
145	CHANGING PAPER, RIBBONS & TAPE IN TELEPRINTERS (COMM)	.39

Comments

In general the tasks at the high end of the scale are those, that if not done correctly, could lead to a mission not being completed successfully. Tasks that are "critical" in this regard are not necessarily difficult, nor do they necessarily demand a great deal of concentration or mental effort as evidenced by the fact that vectors relating most closely to these properties in the ROS configuration have respective cosines of .16, .18, and .24 with Dimension IV. (This is to be expected, since a simple task like throwing an integral switch could be critical).

Dimension IV in the ROS configuration is more clearly interpreted as relating to "criticality", particularly when relationships to the property vectors are considered (see Table 29). While Dimension IV in the OBS1 and OBS2 configurations are strongly related, the ROS Dimension IV does not compare as closely to the corresponding dimension in any of the other groups.

Table 16 (continued)

Task Number	Task Statement	Scale Value
	Not so Crucial to Successful Completion of Mission (Larger Error Tolerance)	
057	ORIENTING MAD (ID-378)	-.43
041	INSPECTING ASH EQUIPMENT IN NOSE ON PPI (DETECTION)	-.43
044	CHECKING SERVICEABILITY OF ASH SYSTEM (DETECTION)	-.43
039	IDENTIFYING & RECTIFYING ASV FAULTS LISTED IN CHECKLIST (ASV)	-.44
048	IDENTIFYING A "SIGNAL OUT" SITUATION ON ASH (DETECTION)	-.44
038	COMMUNICATING WITH PILOT IN CALM CONFIDENT MANNER DURING WX (ASV)	-.55
047	IDENTIFYING A PEAK ON AN ASH TRADE (DETECTION)	-.46
055	CENTERING RECORDER PEN USING OUTPUT BALANCE & PEN POSITION CONTROLS-MAD	-.47
054	SETTING PEN SELECTION SWITCHES FOR MAD ON RECORDER (DETECTION)	-.55
030	INSPECTING SCOPE FOR SIMILAR CONTOURS WHEN MAP READING (ASV)	-.57

Table 17

AVE Dimension V: Source Initiating Activity

Task Number	Task Statement	Scale Value
	Internal or Self Initiated	
120	PFI & PRESETTING OF ECM COMPONENTS	.65
162	KEYING ORESTES (COMM)	.63
160	LOOKING UP A MANUAL FREQUENCY ON ARC38 (COMM)	.62
159	SETTING A PRESET FREQUENCY ON ARC38 (COMM)	.61
161	SETTING A MANUAL FREQUENCY ON ARC38 (COMM)	.61
152	CHECKING SERVICEABILITY OF ARC505 WITH EXTERNAL AGENCY ON VOICE (COMM)	.56
146	LOADING & CHECKING LP BLACK BOX (COMM)	.56
149	CALIBRATING MODEM (COMM)	.51
150	CHECKING SERVICEABILITY OF ORESTES CONTROL BOX (COMM)	.51
043	SETTING PEN SELECTION SWITCHES FOR ASH ON RECORDER (DETECTION)	.51

Comments

Tasks at the high end of this scale tends to be those which the individual himself initiates. Those at the opposite end tend to be initiated by others, either within or outside the aircraft.

	Externally Initiated	
015	ALIGNING RANGE & BEARING MARKERS ON TARGET (ASV)	-.48
083	CHECKING & CALLING BUOY SERVICEABILITY (RX/AUDIO/HYDROPHONE) (JULIE)	-.49
062	SETTING SWITCHES FOR BUOY SELECTION (SSQ47-DETECTION)	-.50
063	DETECTING TARGET FROM RECORDER (SSQ47)	-.51
086	LOADING TAPES IN AR102 TAPE RECORDERS (DETECTION)	-.52
079	MEASURE/PASS SINGLE ECHO MASTER RANGES (JULIE-DETECTION)	-.54
080	MEASURE/PASS SINGLE ECHO SLAVE RANGES (JULIE-DETECTION)	-.56
082	MEASURE & PASS JULIE DOUBLE ECHO RANGES-DROPPED SIMULTANEOUS, DEEP/SHALLOW	-.62
074	CHANGINE PAPER IN AJH501 RECORDER (JULIE-DETECTION)	-.63
081	MEASURE/PASS DOUBLE ECHO RANGES (JULIE-DETECTION)	-.74

Table 18

AVE Dimension VI: Teamwork or Cooperation Involved

Tasks Done Primarily by Self

Task Number	Task Statement	Scale Value
063	DETECTING TARGET FROM RECORDER (SSQ47)	.63
019	ADJUSTING SCOPE OPTIMAL TARGET PRESENTATION (ASV)	.61
141	ADJUSTING AE GAIN AND/OR ATTENUATION DURING HOMING (ECM)	.60
164	PRESSING SYNCH BUTTON ON TIME CHECK CUE ON LF (COMM)	.58
090	CHECKING & RECORDING ALL ASW STORES ON BOARD DURING PFI (RETRO)	.58
169	IDENTIFYING & RECTIFYING COMM EQUIPMENT FAULTS LISTED IN CHECKLIST	.57
002	CHECKING METER VOLTAGES ON TURN-ON (ASV)	.57
062	SETTING SWITCHES FOR BUOY SELECTION (SSQ47-DETECTION)	.56
130	ADJUSTING SCOPE FOR BEST D/F SIGNAL (ECM)	.55
066	TAKING TARGET RANGE CHECK WITH STOP WATCH (SSQ47-DETECTION)	.50

Comments

Loading on the high end of this dimension are tasks the individual does primarily by himself. Further, those tasks tend to require concentration (.59), are difficult (.49), are important (.38), may have to be done quickly (.45), and require considerable mental effort (.67), as evidenced by the direction cosines between this dimension and the configuration vectors relating most strongly to these properties. Tasks at the opposite pole tend to demand manual skill (-.49) and cooperation (-.50). These relationships represent a profile that is consistent with the interpretation given this scale.

Many of the tasks loading at the low end of the dimension require cooperation in the sense that the individual must rely on someone else to do, or not do something (i.e., throw a switch, before or while the task is performed).

Tasks Requiring Reliance on Someone Else to Do or Not Do Something

108	DROPPING PARACHUTE FLARE (ASW)	-.37
094	UNLOADING & TURNING OFF RETRO (ASW)	-.41
157	LOGGING ALL MESSAGES RECEIVED & TRANSMITTED ON LF (COMM)	-.42
104	UNLOADING LIBRASCOPE (ASW)	-.44
099	UNLOADING SONO CHUTES (ASW)	-.46
097	LOADING SONOS IN CHUTES (ASW)	-.48
078	IDENTIFYING & CALLING ECHOES DURING JULIE OPERATIONS (DETECTION)	-.53
077	SELECTING PROPER BUOYS DURING JULIE OPERATIONS (DETECTION)	-.56
140	ESTIMATING RELATIVE TARGET MOVEMENT & DRIFT (ECM)	-.59
116	INSTALLING MAI CHUTES	-.67

Table 19

AVE Dimension VII: Discreteness of Task Event

Tasks Involving Single Discrete Behaviours

Task Number	Task Statement	Scale Value
120	PFI & PRESETTING OF ECM COMPONENTS	.71
093	FIRING RETRO LOCALLY (ASW)	.68
109	RETURNING PARACHUTE FLARE TO STORAGE (ASW)	.59
052	CHECKING BATTERY IN MAD SET (DETECTION)	.54
104	UNLOADING LIBRASCOPE (ASW)	.53
108	DROPPING PARACHUTE FLARE (ASW)	.52
094	UNLOADING & TURNING OFF RETRO (ASW)	.51
110	CHECKING HULCHER CAMERA SERVICEABILITY (FRAME COUNTER & MOTOR VIBRATION)	.51
107	PREPARING PARACHUTE FLARE FOR DROP (ASW)	.50
059	IDENTIFYING & CALLING MAD MARKS (DETECTION)	.46

Comments

Tasks loading high on this dimension tend to be those requiring relatively discrete action. It appears to be coincidental that many of these tasks are also of a heavy physical nature..

Tasks on the low end of this scale appear to require chained sequences of events that can take place over reasonably long time periods. Many of the activities involved in these tasks are such that a later step is contingent on what occurs in a former one. That is, step X generally will not be performed until step X-1 has been completed.

Tasks Involving Chained and Sequenced Activities

Task Number	Task Statement	Scale Value
138	IDENTIFYING & RECTIFYING ECM FAULTS LISTED IN CHECKLIST (ECM)	-.38
065	CATEGORIZING DOPPLER (SSQ47-DETECTION)	-.39
165	SELECTING BEST FREQUENCY FOR USE WITH AGENCY (COMM)	-.39
147	LOADING & CHECKING HF BLACK BOX (COMM)	-.39
134	CHANGING TUNING UNITS & RF CALIBRATING (ECM)	-.40
123	CHECKING XTAL SERVICEABILITY OVER TUNERS RANGE (ECM)	-.41
007	COMPUTING DRIFT FROM DRIFT & HEADING MARKERS (ASV)	-.41
148	CHECKING SERVICEABILITY OF ARO 38 WITH EXTERNAL AGENCY (COMM)	-.42
140	ESTIMATING RELATIVE TARGET MOVEMENT & DRIFT (ECM)	-.47
156	LOGGING ALL MESSAGES RECEIVED & TRANSMITTED ON HF (COMM)	-.58

Table 20

AVE Dimension VIII: Danger Level in Task Event

Task Number	Task Statement	Scale Value
	Hazardous Situation	
118	DROPPING MAI BAGS IN SEQUENCE	.56
105	PREPARING MARKER MARINE FOR LAUNCH (ASW)	.55
100	KEEPING CHUTES LOADED IAW NAVS INSTRUCTIONS (ASW)	.55
095	CHECKING SONO CHUTES ON PFI (ASW)	.54
098	FIRING SONO CHUTES LOCALLY (ASW)	.52
107	PREPARING PARACHUTE FLARE FOR DROP (ASW)	.52
103	FIRING LIBRASCOPE MANUALLY (ASW)	.51
109	RETURNING PARACHUTE FLARE TO STORAGE (ASW)	.51
092	CLEARING JAMMED RETRO (ASW)	.50
099	UNLOADING SONO CHUTES (ASW)	.50

Comments

Tasks at the high end of this dimension are either hazardous in and of themselves or are performed under hazardous conditions. Tasks at the lower end tend to involve rather safe activities usually performed under safe conditions.

Configuration vectors corresponding to the properties concentration and speed show direction cosines of .38 and .34 with this dimension. These relationships are logical since handling these dangerous tasks safely requires considerable concentration, and many (but not all) of the hazardous tasks are performed when quick response is of the essence.

	Nonhazardous Situation	
127	CHECKING KD2 CAMERA (ECM)	-.44
150	CHECKING SERVICEABILITY OF ORESTES CONTROL BOX (COMM)	-.46
112	SETTING UP SHUTTER SPEED & LENS OPENING ON HULCHER (ASW)	-.56
130	ADJUSTING SCOPE FOR BEST D/F SIGNAL (ECM)	-.61
062	SETTING SWITCHES FOR BUOY SELECTION (SS047-DETECTION)	-.62
019	ADJUSTING SCOPE OPTIMAL TARGET PRESENTATION (ASV)	-.63
141	ADJUSTING AE GAIN AND/OR ATTENUATION DURING HOMING (ECM)	-.64
090	CHECKING & RECORDING ALL ASW STORES ON BOARD DURING PFI (RETRO)	-.64
169	IDENTIFYING & RECTIFYING COMM EQUIPMENT FAULTS LISTED IN CHECKLIST	-.64
164	PRESSING SYNCH BUTTON ON TIME CHECK CUE ON LF (COMM)	-.66

Table 21

AVE Dimension IX: Degree of Imagery or Orientation to Earth

Task Number	Task Statement	Scale Value
	Activity or Imagery (i.e., Keeping HC Log) Related to Earth's Surface	
116	INSTALLING MA1 CHUTES	1.03
078	IDENTIFYING & CALLING ECHOES DURING JULIE OPERATIONS (DETECTION)	.78
077	SELECTING PROPER BUOYS DURING JULIE OPERATIONS (DETECTION)	.77
140	ESTIMATING RELATIVE TARGET MOVEMENT & DRIFT (ECM)	.70
112	SETTING UP SHUTTER SPEED & LENS OPENING ON HULCHER (ASW)	.53
114	TAKING PICTURES IN NOSE WITH HULCHER CAMERA	.51
115	KEEPING ACCURATE HULCHER CAMERA LOG (ASW)	.48
008	SETTING UP GIVEN SECTOR FOR TRANSMITTER CHECK (ASV)	.41
091	TURNING ON AND LOADING RETRO (ASW)	.39
111	ASSESSING WX FOR PROPER HULCHER CAMERA SET-UP (ASW)	.39

Comments

The profile of task scale values on this scale tends to lead one to feel the dimension differentiates tasks on the basis of whether they are surface oriented or not (either in terms of direct activity or imagery to support activity). Tasks having to do with the earth's surface (ground / water) tend to load on the high end.

Scale value variability on this scale is not as large as in many of the others, particularly at the low end. This might suggest that the low end of the scale is not well defined.

Task Number	Task Statement	Scale Value
	Activity not Oriented to Earth's Surface	
120	PFI & PRESETTING OF ECM COMPONENTS	-.24
068	SETTING SWITCHES AND KNOBS FOR OPERATION (AURAL LISTENING-DETECTION)	-.25
011	PERFORMING POST TAKE OFF CHECK (ASV)	-.25
148	CHECKING SERVICEABILITY OF ARC 38 WITH EXTERNAL AGENCY (COMM)	-.25
125	SETTING UP PANORAMIC PRESENTATION ON SCOPE (GRASS-ECM)	-.28
157	LOGGING ALL MESSAGES RECEIVED & TRANSMITTED ON LF (COMM)	-.29
168	REPLYING AS REQUIRED TO EXTERNAL AGENCIES IN COMM OPERATIONS	-.30
006	CHECKING & ALIGNING HEADING MARKER (ASV)	-.32
085	INSPECTING AR102 TAPE RECORDERS DURING PFI (DETECTION)	-.33
005	CHECKING ASV SECTOR SCAN (ASV)	-.34

Table 22

AVE Dimension X: Housekeeping Functions

Task Number	Task Statement	Scale Value
	Primarily Checking and Housekeeping Functions in Which Unsuccessful Performance Not Generally Crucial to Mission Success	
115	KEEPING ACCURATE HULCHER CAMERA LOG (ASW)	.97
114	TAKING PICTURES IN NOSE WITH HULCHER CAMERA	.93
112	SETTING UP SHUTTER SPEED AND LENS OPENING ON HULCHER (ASW)	.59
083	CHECKING & CALLING BUOY SERVICEABILITY (RX/AUDIO/ HYDROPHONE) (JULIE)	.51
120	PFI & PRESETTING OF ECM COMPONENTS	.50
084	IDENTIFY & RECTIFY JULIE FAULTS LISTED IN CHECKLIST (JULIE DETECTION)	.45
089	IDENTIFYING & RECTIFYING AR102 FAULTS LISTED IN CHECKLIST (DETECTION)	.45
088	SETTING UP AR102 REMOTE CONTROL TO RECORD A FACILITY (DETECTION)	.43
085	INSPECTING AR102 TAPE RECORDERS DURING PFI (DETECTION)	.42
076	CHECKING CHART SPEED WITH ROOF TOP CHECKER (JULIE-DETECTION)	.40

Comments

To some extent, tasks loading on both ends of this dimension involve checking and housekeeping functions. The difference between them appears to be the fact that those on the low end, if not done correctly, by themselves are more likely to be responsible for mission failure. The nature of the profile of direction cosines between this dimension and configuration vectors corresponding most closely to the properties concentration (-.38), importance (-.69), cooperation (.33), and speed (-.49) reinforces this interpretation.

	Primarily Checking and Set up Functions that, if Done Incorrectly, Could Lead to Mission Failure	
004	CHECKING & SETTING ASV TILT (ASV)	-.34
096	SETTING SWITCHES ON SONOS FOR PROPER DEPTH/LIFE (ASW)	-.35
165	SELECTING BEST FREQUENCY FOR USE WITH AGENCY (COMM)	-.35
121	CHANGING ECM ANTENNA IN AFT LOWER FUSELAGE COMPARTMENT (ECM)	-.36
153	SETTING ARC505, TELEPRINTER, & I/C SWITCHES TO TRANSMIT A RATT MESSAGE	-.36
064	DETECTING TARGET FROM HEADSET (SSQ47)	-.36
024	MAINTAINING DRIFT MARKER ON TARGET (ASV)	-.39
131	CENTERING SIGNAL ON PAN PRESENTATION (ECM)	-.39
001	PFI & PRESETTING ASV21 COMPONENTS (ASV)	-.48
091	TURNING ON AND LOADING RETRO (ASW)	-.51

Obstacles to and Incentives for Standardization of Task Analysis Procedures

by

Robert W. Stephenson
and
Hendrick W. Ruck

Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper are those of the authors and are not necessarily those of the United States Air Force.

A number of critical papers have been written regarding the status of task analysis in the Air Force and the assumptions upon which task analysis is managed within the Department of Defense. Montemerlo and Harris (1978) cited a long list of such papers at the 1978 Annual Convention of the American Psychological Association. One of the major conclusions in their paper is:

> "...while everyone agrees on the need for task analysis, there is almost no agreement as to what it is."

The writers go on to conclude that "the procedural approach to task analysis has not and cannot work," because task analysis is essentially a judgmental process. Many other experts referenced in their paper had come to similar conclusions.

In the face of so many opinions that task analysis should not be proceduralized in the first place, one feels a bit awkward presenting a paper about standardization of task analysis procedures. As shall be seen, however, our own position is not incompatible with that presented by Montemerlo and Harris.

Different Kinds of Task Analysis

Before discussing obstacles and incentives for standardization, it is necessary to clarify what kind of task analysis we are talking about. In the Air Force, one can distinguish six different kinds of task analysis as shown in Table 1: the task analysis associated with the design of new weapons, which is an intrinsic part of the research and development process; the task analysis associated with the preparation of Technical Orders after the weapon system has been designed (these first two types of task analysis are usually conducted by a weapons development contractor during the weapons development process);

Table 1 Six Different Kinds of Task Analysis

Task Analysis For...	Objective	Requirements for Judgment	Acceptable Personnel
New systems design	Specify task procedures for undesigned weapon systems	Extremely High	Developmental systems design engineers
Systems documentation	Specify task procedures for previously designed weapon systems	Very High	Systems design engineers
Systems evaluation	Evaluation of technical orders by government experts	High	Systems analysts
Training design for unperformed jobs	Training performance objectives and design of new courses for new jobs	High	Professionally trained ISD analysts
Training design for existing jobs	Training performance objectives and design of courses for existing jobs	Moderate	Subject matter experts with expertise in training
Course revision	Minor additions to an existing course	Moderate	Subject matter experts who take short courses for orientation

the task analysis that is conducted when the contractor's Technical Orders are evaluated by the government (in the Air Force, these evaluations are typically conducted at Edwards Air Force Base); the task analysis that is conducted after the jobs have been established but before occupational survey data are available; the task analysis that is conducted after occupational survey data are available but before the course has been written; and finally, the task analysis that is conducted in order to revise an existing course.

We agree with the various experts cited in Montemerlo and Harris (1978) about the need for judgment and experience--as opposed to standardized procedures--especially for the three or four kinds of task analysis that appear early on this list. We would maintain, however, that both the feasibility and desirability of standardization increases as you get further and further away from the original weapons development process. One of the reasons that standardization is desirable is that the personnel who conduct the task analysis in these later stages are typically not professional analysts. There are some hard realities in the Department of Defense budget that force us to use enlisted subject matter experts who are not professionally trained as system analysts or educators. To the extent that such personnel are used--the need for standardization and simplication of procedures increases.

Failure to make distinctions between these various types of task analysis can lead to a lot of confusion. It is not unusual for someone to attack task analysis procedures designed for revising courses because they are not documented with the kind of detail required for weapons systems design. ISD experts sometimes get upset, for example, because all of the procedures designed for brand new jobs are not being used in the revision of Air Force courses. The probability of this reaction is increased by the fact that most ISD manuals are designed for new courses rather than the revision of existing ones. We have also encountered a similar type of confusion in which subject matter experts are asked to do jobs that they are not qualified to do, because it is assumed that a person who can do one type of task analysis is qualified to do other types that really require professional training.

Granted that some kinds of task analysis do require expert judgment by professionally trained personnel and some do not, the question addressed in this paper is, "To what extent should we standardize procedures for those kinds of task analysis that are currently being conducted by subject matter experts rather than by professional analysts?" In other words, to what extent should we standardize procedures for the two types of task analysis shown in the last two rows of Table 1.

Incentives for Standardization

There are many incentives for standardization, and most of them are relatively obvious (see Table 2). One can avoid duplication of effort, facilitate communication, improve the amount of management control, provide a consistent basis for evaluation, help inexperienced subject matter experts to benefit from the experience of professional analysts, and so on. These things are especially important in the military environment, where there is rapid turnover of key personnel and rapid technological change in the jobs.

Table 2 Incentives for Standardization

- Minimize duplication of effort
- Facilitate communication
- Improve management control
- Provide a consistent basis for evaluation
- Provide inexperienced personnel with useful guidelines
- Facilitate training of new task analysts

Obstacles to Standardization

The obstacles to standardization are perhaps not so obvious (see Table 3). First, let us deal with the obstacles to DOD-wide standardization, across Army, Navy, Air Force, and Marine Corps. In the first place, the jobs are different. At one extreme, the Navy must provide personnel with a wide diversity of qualifications for assignments to small ships. The Navy consequently has very broadly defined job categories, called ratings. At the other extreme, the Air Force, which typically has large installations that work with highly specialized equipment, has very technical jobs in specific job categories that are more narrowly defined than the Navy ratings. The Army and Marine Corps fall in between the extremes.

Another obstacle to standardization is that the various task analysis organizations are staffed differently. The Navy seems to have extremely qualified people at its new Instructional Program Development Centers. The Air Force, by contrast, has had to react to budget cuts by repeatedly decreasing the number of professionally trained personnel at the Air Force technical training centers. The Marine Corps has the fewest professionals while the Army is probably more similar to the Air Force than it is to the Navy.

Another obstacle to DOD-wide standardization is that the occupational survey inputs to ISD personnel regarding established jobs varies from service to service. All services do use occupational survey data, but the extent to which these data are analyzed before they are submitted for use in task analysis and training program design varies considerably from service to service. The Air Force, which originally designed the occupational survey methods used today (Christal, 1974),

Table 3 Obstacles to Standardization of Task Analysis Procedures

Obstacles to DOD Wide Standardization

- Jobs and job categories are different
- Qualifications of ISD and task analysis staffs are different
- Inputs from occupational survey data are different
- User orientation of occupational measurement centers is different

Obstacles to Standardization Within a Single Service

- Different requirements for task analysis associated with combat crew training
- Different resources and professional expertise

Unresolved Issues Regarding the Design of a Task Analysis Manual

- Task analysis for training versus task analysis for multiple users
- Task analysis documentation that is of marginal utility

has taken a great deal of interest in occupational survey data, and the data that are provided to Air Force training centers are rapidly growing more sophisticated and better organized. The data provided by the Air Force are also much more detailed than that provided by the other services. This is partly a function of the way in which the jobs are defined. If the job categories are relatively specific, as is true of Air Force career ladders, it is possible for the occupational survey information at the task level to also be specific. The Navy, since it uses broadly defined job categories, almost necessarily uses task statements that are broadly defined. If they don't do so, the task inventories will be too long, and there will be problems with the quality of the data.

Other differences between the services are associated with the way in which the occupational measurement centers are organized. In the Air Force, the Occupational Measurement Center is part of the Air Training Command, and training applications are given extremely high priority. In the other services, the occupational measurement center is part of a Military Personnel Center, and other uses of occupational survey data (e.g., classification, job satisfaction studies, etc.) have high priority, while training applications seem to have less priority.

Another set of obstacles to standardization exists within each service. For example, in the Air Force task analysis is conducted at

216

Combat Crew Training Schools (each of which is associated with a major command) as well as at Air Force-wide Technical Training Centers (which are part of Air Training Command). These two parts of the Air Force typically follow different task analysis procedures, and typically approach the problem in different ways. The Combat Crew Training Schools (CCTSs) conduct a very sophisticated kind of task analysis since they must deal with teams of personnel rather than individuals. These teams of personnel, moreover, are involved in complex combat scenarios with multiple weapon units and multiple delivery systems. The needs of the two types of Air Force organizations are so different that there are two sets of complaints about the Interservice ISD manual (Interservice Committee for Instructional Systems Development, 1975). Some of the ISD personnel at the CCTSs complain that these procedures are too simple. The ISD personnel at the Technical Training Centers complain that the same interservice procedures are too complex.

The need for additional complexity in CCTSs is illustrated by a case in which the tasks associated with a combat attack plane were analyzed three times. First the standard interservice ISD procedures were used. Unfortunately, these interservice task analysis procedures were deemed inadequate because there was not enough emphasis upon performance objectives. The whole task analysis was redone using Mager's performance objectives (Mager and Pipe, 1976) which seemed to help, but this, too, was not satisfactory. The task analysis was redone again using combat team descriptions of the tasks as part of complex combat scenarios. The reaction to the interservice manual is exactly the opposite at ATC technical training centers where it is simply considered too complex. At ATC schools, the interservice manual is primarily used for guidance in the design of more simplied procedures for local use.

The amount of resources and expertise available for task analysis also differ sharply between Technical Training Centers and Combat Crew Training Schools. The CCTSs often use contractors, whereas the Technical Training Centers tend to use military subject matter experts in each specialty. The qualifications of the staff assembled by a contractor organization tend to be of very good quality. They also tend to be very expensive.

Another set of obstacles exists because of unresolved issues regarding the design of a task analysis manual. To what extent should a training task analysis provide information for multiple users of task analysis information? To what extent, for example, should the performance objectives designed for training purposes be useful to the people who establish performance objectives for promotion? Another important unresolved issue is the question of how much documentation of task analysis procedures is really needed. Suppose that you have in your hand a detailed Plan of Instruction (POI) containing behavioral objectives for each block of instruction. Do you still need a lot of task analysis documentation to back up that POI, or can it be argued that the end product is all that is really required? If one is in a military training command, it is easy to argue that the task analysis documentation for purposes other than training is not needed, or--if it is needed--that it should come out of somebody else's budget rather than your own. Consideration must also be given to documentation that is of marginal utility. Granted that some documentation is essential, one can argue that the value provided by additional amounts of documentation is less and less until--eventually--the additional documentation seems to be more trouble than it is worth.

The Case for Non-Standardized Task Analysis Procedures

As long as the various military services are staffed differently, organized differently, have different kinds of input information, and differing amounts of professional expertise and differing customer orientations, a very good case can be made for permitting each service to have its own task analysis procedures. This does not mean that the various services cannot derive mutual benefit from improvements in task analysis methodology or standardized formats for shared information. In our current work on the design of an Air Force task analysis manual, we have already contacted people from Air Force Combat Crew Training Centers as well as people in Army, Navy, Marine Corps, and Coast Guard. They will be invited to help themselves to any of our methods that look useful to them. We are also open to suggestions for standardized formats for sharing information about the results of task analysis efforts.

The Case for a Standardized Task Analysis Data Bank

One can make a much stronger case for a standardized task analysis data bank than we can for a task analysis manual (see Table 4). There are many jobs in the military services (e.g., plumbers, carpenters, machinists) that are so similar that it would be a tremendous waste of effort if all services were to conduct independent task analyses of their own. Yet that is exactly what has happened and is currently happening at this time. The job of carpenter, for example, is analyzed by all military services. This is certainly not the way in which task analyses are handled at the Vocational Technical Education Consortium (VTEC) of Southern States (Hirst, 1975). In this consortium, the task analysis work is divided so that each state only does its proportionate share of the task analysis work in areas of general interest. Georgia, for example, may analyze the job of carpenter and Mississippi may analyze the job of plumber. There are also many advantages involved in having analysis information available on computer. For example, the computer can generate field survey sheets that can be used for validation studies of the task analysis worksheets.

There are problems with standardized data bank proposals, however. While it is true that a task analysis data bank is highly cost effective if one is starting out from scratch to develop task analysis information

218

Table 4 The Case for a DOD-wide Computerized Task Analysis Data Bank

Incentives

Minimize duplication of effort
Document previous efforts that have not been well documented
Generate work center catalogs for OJT
Rapid updating
Facilitate sharing of information
Computer assembled forms for field surveys

Obstacles

Uncertainty as to whether documentation is really needed
Uncertainty as to cost effectiveness
Prior need for agreement on a standardized format for information to be shared

for all jobs in the Department of Defense, that is not the situation in which we find ourselves. Almost all DOD jobs have already been analyzed, and Plans of Instruction with behavioral objectives already exist. Under such circumstances, a DOD-wide task analysis data bank is not a way of avoiding work, it is a requirement to do work that would not otherwise be accomplished at all. One could conceivably argue that undocumented work is work of poor quality, and that a standardized task analysis data bank should be required in order to upgrade the quality of the information. We don't really know whether this is true or not, however, since we do not have acceptable measures of the task analysis information that is presently on file, nor do we have good information about the cost advantages of redoing the work if it is of poor quality.

One thing is certain--if one already has end products in the forms of behavioral objectives for Plans of Instruction (POIs)--it is very difficult to convince training executives that they should undertake a massive documentation effort for task analysis data. The training commands already have the POIs that such an effort would produce and it is the POIs--especially POIs that are provided with behavioral objectives --in which they are most interested.

We conclude that the disadvantages of standardizing task analysis procedures outweigh the advantages (see Table 5). This does not mean that all forms of standardization are undesirable. We at HRL, for example, are considering plans for a task analysis data bank for critical tasks that are scheduled for On-the-Job Training (OJT). The objective is to provide each work center with a catalog containing information about the performance standards, the steps to be followed in accomplishing the task, the relevant Technical Order references, and so on--for all the critical tasks in a particular work center. In developing this task analysis data bank, we do not, however, propose to re-do the existing task analyses for every job in the Air Force. The Air Force cannot afford to do such a thing even if we wanted it to--which we do not.

Granted that we should not attempt to establish a task analysis data bank all at once, it is still possible to establish such a data bank over a period of ten or twenty years by standardizing all new task analysis documentation. This focus upon new documentation would still permit us to divide up the responsibilities for documentation among the various military services, so as to minimize duplication of

---

Table 5 Conclusions

Question	Answer	Comment
Should task analysis procedures be standardized?	No	Current differences in jobs, job categories, staff qualifications occupational survey inputs, and user orientation are too great.
Should information about task analysis procedures be shared?	Yes	Communications are excellent in this respect.
Should task analysis data for all courses be redone in a standardized format?	No	Even though documentation may be of poor quality and inconsistent from service to service, the task analyses have already been conducted and courses have already been designed. Work would have to be redone.
Should a task analysis data bank be generated to facilitate OJT in critical tasks?	Yes	Plans to provide computer-assembled work center catalogs containing task analysis data for use in OJT are currently being prepared.
Should standardized output be required when new task analysis efforts are documented?	Yes	An interservice group of representatives should be asked to consider procedures for sharing new task analysis data.

effort. It is a safe assumption that one would--ten or twenty years from now--have documentation that would be of much higher quality than it is today. An interservice group of advisors should be asked to consider this possibility at a conference to be scheduled early next year.

As mentioned earlier in this paper, many of these questions require information about the importance of good quality documentation. The need for documentation is complicated by the fact that many opportunities exist for quick fixes downstream. Since many other procedures can also improve the quality of training, how important is the initial task analysis? We know, for example, that effective use of occupational survey and task training emphasis data can keep us from overtraining--we also know that complaints from the field can keep us from undertraining. So we have two feedback loops that will gradually improve our courses over a period of several years regardless of the quality of the initial task analysis.

Those who recommend standardization of documentation would be completely correct if every service were starting out fresh to conduct task analyses for every occupation. But they are not. On the contrary, the typical task analysis requirement nowadays involves a minor scrubdown of an existing course that is already well defined in terms of behavioral objectives. The task analysis documentation may be non-detailed or even non-existent--but we don't really know how important a lack of documentation really is.

One of the reasons that large quantities of documentation seem so attractive to many people is that they tend to think in terms of task analysis for systems design or task analysis for new courses. People tend to assume that if a large amount of documentation is good for the human eningeers and the systems designers, then it must be equally good for the trainers. In actual fact, a similar amount of documentation may or may not be needed for the trainers, but we should at least recognize the price tag. If it is needed--we are going to have to reaccomplish hundreds of manyears of work for which the responsible training organizations do not have adequate resources. This can be redone all at once, with a lot of duplication of effort--or it can be redone gradually over a period of many years as part of the normal updating function.

The only way to really resolve this issue is to collect systematic evidence regarding the cost effectiveness of standardized documentation. We can certainly support the need for this documentation without equivocation. However, until the evidence has been collected and the cost determinations have been made, proposals for high priority standardization of task analysis documentation are more than just a little bit late. They are proposals for large expenditures of time and effort without any systematic evidence that these expenditures are really worthwhile.

References

Christal, R.E. The United States Air Force occupational research project. AFHRL-TR-73-75, AD-774-574. Lackland AFB, TX: Occupational Research Division, January 1974.

Hirst, B.S., Jr. The instructional systems model of the Vocational-Technical Education Consortium of States used to develop performance objectives, criterion-referenced measures and performance guides for learners. In P.E. Schroeder (Ed.), Proceedings of a symposium on task analysis/task inventories (UN Series No. 10). Columbus: The Ohio State University, The Center for Vocational Education, 1975.

Interservice Committee for Instructional Systems Development. Interservice Procedures for Instructional Systems Development (NAVEDTRA 106A, five volumes). Fort Benning, GA: U.S. Army Combat Arms Training Board, August 1975.

Mager, Robert F. and Pipe, Peter. Criterion referenced instruction: Analysis, design and implementation, Participant Manuals (Revised Edition). Los Altos Hills, CA: Mager Associates, Inc., 1976.

Montemerlo, Melvin D. and Harris, Ward A. Angels, pinheads, and task analysis. Paper presented at the 1978 Annual Convention of the American Psychological Association, Toronto, Canada, Sep 1, 1978.

222

TASK ANALYSIS: DESTINATION OR JOURNEY

Dr. Melvin D. Montemerlo
Dr. Frank M. Aversano
U. S. Army Training Support Center
Ft. Eustis, VA 23604

The Systems Approach to Training (SAT), as it was known in the 1960's, or Instructional Systems Development (ISD), as it is now known, has become the pre-eminent concept of modern instructional technology. More than 100 SAT/ISD manuals have been published during the last quarter century (Montemerlo and Tennyson, 1976). Each of these manuals breaks down the process of course development into a linear sequence of steps designed to be carried out by laymen, that is, by personnel with little or no background in instructional design (Klein, 1977). Although the manuals differ as to what the steps are and how they are to be accomplished, they all agree on two things: (1) that one of the steps is "Task Analysis", and (2) that task analysis is proceduralizable (that is, that it can be reduced to a well-defined, pre-stated sequence of actions). Most SAT/ISD manuals present task analysis as nothing more than the filling out of a pre-designed form on each of the tasks to be trained.

The SAT/ISD view that task analysis is a routinized procedure has given rise to the misperception that following the procedure will surely lead to TRUTH (i.e., to a well-defined result which is the best definition of the tasks to be taught, and which all task analysts would agree with). Unfortunately, task analysis has come to be viewed as a destination, and not as a journey. It is the hypothesis of this paper that the latter analogy is more appropriate. In any given task analysis, the distance the journey progresses depends on: the time and money available; the experience, skill, rank, personality and political acumen of the task analyst (not to mention his knowledge of the subject matter area); the type of subject matter involved; the existance of similar analyses; and the co-operativeness of the people being analyzed.

Task analysis is a purely rational process (in most cases). Philosophy is a purely rational process (in most cases). The scientific method came about as a recognition of the limitation of purely rational processes. At first blush, it would seem that heavier rocks would fall faster than light rocks. After all, light rocks fall faster than feathers. The scientific movement didn't really throw out the armchair; it merely asked the person sitting in it to get up after he finished thinking, and test out his conclusion. Task analysis has been pretty much a 100% return to the armchair, for in almost all training development projects there is little enough money to do a purely rational task analysis, never mind going on and testing these ideas. Even if there were the time and money to test the task analysis, what would it be tested against?

For any job, consider the set of all possible task analyses of that job. One couldn't know if the best possible task analysis was in hand until the entire universe of task analyses on that job was done. But even then, upon what variables would you judge which was best, or even which ones were adequate? Since the only real measure of the goodness of a task analysis is the degree of its usefulness in generating the training program, one would have to use each of the task analyses to develop a training program and then assess the relative goodness of the resulting courses. In case of a tie, the analysis which was the easier to use would be the winner.

Any one who doubts that there would be much variation in task analyses which are independently done on the same task should do a little experiment. Select a very simple task, and have any two people analyze it independently. (Three, four or five people are even better.) While in his doctoral program at Penn State, Dr. Montemerlo participated in such an exercise. One Friday, one of his professors (in an instructional technology course) asked the five students to do a task analysis of a relatively straightforward mathematical task. The five returned on Monday with documents ranging in length from a few pages to a small volume. The professor asked the students to peruse the analyses done by each of the other students. He then stated that they could dispose of their analyses; he didn't need to see them. The dismayed students were given the following explanation: "I asked you to do a task analysis. I gave you the task, but I never told you the purpose your task analysis was to serve. Without that information you can't do a meaningful task analysis." His message was clear and his medium was effective. However, even if he had told the students the goal of the analysis, there would have been large differences in the result.

Doing a task analysis is much like taking a projective test. There is monumental room for variance in the results. One rationale for this is provided by John Holt (1976) in his latest book "Instead of Education." He states,

> "It may be true, at the level of words, to say that anyone doing a difficult thing well is using a variety of skills. But this does not mean that the best way to teach a difficult act is to break it down to as many separate skills as possible and teach them one by one."

Holt is pointing out the artificiality that necessarily exists whenever human performance, which is continuous, is sliced into discrete "tasks."

Holt's hypothesis is not new. He attributes it to Alfred North Whitehead. Another great educator who espoused it is R. B. Miller, the father of modern task analysis. Miller (1966) stated,

> "A task is a fairly arbitrarily bounded set of activities. A rigorous operational definition cannot (and therefore should not) be sought. It is a heuristic term."

At the Air Force's ISD Conference (The Pentagon, 3-5 Feb 76) Burt Cream, of the Air Force's Human Resources Laboratory, described six assumptions of ISD which, he claimed, are unsupportable. All have to do with the task analysis portion of ISD. They are, that:

1) any task can be reduced to a series of stimuli and responses.
2) the resulting task breakdown is the best way to teach the task.
3) the personnel who can do a task best can do the best analysis.
4) a whole task is nothing more than the sum of its parts.
5) defining "successful performance" of a task is straightforward, and,
6) complexity always yields to successful analysis.

Montemerlo (1976) added four more unsupportable assumptions to Cream's list.

7) task analysis is a non-political process.
8) the process requires no creativity.
9) there is one best way to teach any task.
10) one method of analysis is best for all tasks.

The hypothesis that tasks can be broken down into a linear sequence of stimuli and responses has been highly attractive for some time. It was one of the foundations of the behavioral or "S-R" school of psychology. The hypothesis that tasks can't be broken down has also been highly attractive for about the same length of time. It was one of the foundations of the cognitive school of psychology and was popularized by the Gestaltists. Reading up on the history of these two schools may provide some insight into the future of task analysis. In short, the two schools melded. The pure S-R approach was soon given up as untenable and the pure cognitive approach was given up as not very useful.

The S-R people inserted an "O" between the S and the R. The "O" stood for "organism", and was inserted in recognition of the fact that organisms process information coming from the stimulus before reacting to it. The fact that this information processing exists, is obvious, but the fact that it is not open to direct observation is just as obvious. "Intervening variables" and "hypothetical constructs" were hypothesized to account forwhat goes on during this processing. The neo-behaviorists went so far as to hypothesize countless undetectable little s's and r's (called kinesthetic and proprioceptive cues) which occurred between each external stimulis, S, and each response,R. With that development, the behaviorists and the cognitivists came to complete agreement: objective analysis and scientific studies combined could not describe human behavior as a series of stimuli and responses.

The most significant step forward in task analyses in recent years can be found in Klein's 1977 paper entitled, "Phenomenological Approach to Training." He gives a brilliant and forceful argument for recognizing that rational task analysis has its limitations. He finds ISD-type approaches to task analysis suitable for procedural tasks but not for affective skills or for complex perceptual and motor tasks. He finds it useful for describing the relatively choppy performance found at initial stages of learning but not for describing the smooth, highly proficient performance of experts. Klein states:

"Instructor pilots working on ISD teams are frequently charged with developing ISD descriptions of complex performance. They prepare such descriptions, but will typically admit, on an informal basis, that they do not follow those ISD steps while flying."

Klein's major argument is that as a person increases in skill on a task, he experiences shifts in perspective concerning the task. G. A. Miller (1956) had introduced this idea originally under the rubric of "chunking." What is important to the novice is often subsumed into larger chunks of behavior and is no longer consciously thought about. The novice billiards player, for instance, worries only about making his next shot, while the expert is thinking many shots in advance.

Holt (1976) makes a similar point. He feels that analyzing behavior into tasks is artificial in that expert behavior is not a unitary concept. That is, he believes that one does not stop "learning" at some point in time and then start "doing". Great musicians such as Van Cliburn and Earl Scruggs didn't become experts at some point and then level off. To Holt, "learning" is "doing" and both continue until you die. ISD methods generally recognize only two levels of performance, unacceptable and acceptable. Anyone who has ever done a task analysis and attempted to come up with the "acceptable" standards of performance knows the frustration involved and can appreciate Holt's argument.

Instructional technologists in the Navy know that there is an "East Coast Navy" and a "West Coast Navy". Those in the Air Force know that there is a school way of flying and many different operational ways. Those in the Army know that personnel holding the same MOS (Military Occupational Specialty) may have little overlap in the tasks they perform. The bottom line is that the real world of task analysis is much more complex than it appears in the ISD literature.

The recommendation of this paper is not to stop doing task analysis just because it has difficulties, but to realize that it is a journey and not a destination. As with any journey, you should begin a task analysis only after you:

1) know where you intend to go.
2) are willing to pay the price.
3) are prepared for emergencies and changes in venue.
4) have someone along who knows the way.
5) are prepared to stop every so often to assess your progress.
6) realize that someone is waiting for you.
7) are ready to enjoy your trip.

Task analysis may be compared to a specific type of journey--a pioneering, exploratory voyage. You can be sure that:

1) you will be criticized for going.
2) you will be criticized for the route you take.
3) you will find the road rocky at points.
4) there will be dissention in the ranks somewhere along the line.

5) there will be some stressful times and some heartrending decisions.

6) someone who takes the same journey after you will surely take a better route, especially if he has your trip report to work with.

REFERENCES

Holt, John. *Instead of Education*, Toronto: Clarke, Irvin and Co.Ltd., 1976.

Klein, Gary A. *Phenomenological Approach to Training*, Air Force Human Resources Laboratory, Wright Patterson AFB, OH, AFHRL-TR-77-42, Aug 1977.

Miller, G. A. The Magical Number Seven Plus or Minus Two: Some Limits on our Capacity for Processing Information, *The Psychological Review*, 1956, Vol. 63, pp. 81-97.

Miller, R. B. "Task Description and Analysis", In R. M. Gagne (Ed.) *Psychological Principles in System Development*, New York: Holt, Rinehart, and Winston 1966.

Montemerlo, Melvin D. and Tennyson, Michael E. *Instructional Systems Development: Conceptual Analysis and Comprehensive Bibliography.* NAVTRAEQUIPCEN IH-257, Naval Training Equipment Center, Orlando, FL, Feb 1976.

Montemerlo, Melvin D. *Instructional Systems Development: Implications for Further Research*. Paper presented at the Psychology in the Air Force Symposium, Air Force Academy, Colorado Springs, CO, 8-10 Apr 1976.

FOUR FUNDAMENTAL CRITERIA FOR DESCRIBING THE TASKS OF AN OCCUPATIONAL SPECIALTY[1]

Walter E. Driskill, Ph.D.
and
Frank C. Gentner, Capt, USAF

USAF OCCUPATIONAL MEASUREMENT CENTER
OCCUPATIONAL SURVEY BRANCH
LACKLAND AFB, TEXAS 78236

A paper presented at the Military Testing Association Convention

30 October - 3 November 1978

---

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

FOUR FUNDAMENTAL CRITERIA FOR DESCRIBING THE TASKS OF AN OCCUPATIONAL SPECIALTY

Walter E. Driskill, Ph.D.
and
Frank C. Gentner, Capt, USAF

USAF Occupational Measurement Center
Occupational Survey Branch
Lackland AFB, TX 78236

As pressures to lengthen occupational surveys grow, four fundamental criteria for developing task inventories become increasingly important. These essential criteria are (1) each task of the inventory must be time-ratable, (2) each task must communicate in the language of the specialty, (3) each task is mutually exclusive of other tasks in the inventory, and (4) each task must differentiate among workers where actual task performance differs. Besides the communicative, interpersonal, and judgemental skills necessary to elicit job information from career field incumbents, describing the tasks that make up an occupational specialty is a blend and compromise of these four fundamental criteria for task writing.

THE PROBLEM:
LENGTH VERSUS DESCRIPTIVENESS

Theoretically, each occupational specialty should be described at the lowest level of work activity, with activities (tasks) describing a complete and inseparable operation. Often, however, occupational areas are so broad that task description at the lowest level is impractical, because inventory length makes job incumbent response requirements unreasonable. Thus, task inventory development is a matter of compromise between reasonable task list length and the writing of tasks that adhere to the four fundamental criteria.

From the practical viewpoint, a task inventory for an occupational specialty is no more than a sample of the infinite number of activities available for descriptive purposes. The desirable final task inventory captures the essence, or intrinsic nature, of the occupational specialty. It consists of a comprehensive, yet representative set of activities for each subarea of the specialty. Task inventories rarely include all activities comprising an occupational specialty, not only because of the practical constraint of inventory length, but also because of the infinite number of ways that a specialty may be described.

The problem of task list development, then, is to describe the essence of an occupational specialty with a sample of tasks written at the lowest level of specificity consistent with the constraints of length, the fundamental criteria that each task must meet, and the purposes which the ultimate job analysis is to serve. Regarding the purpose of the survey, the results of occupational analysis may be employed for a variety of personnel management purposes. Some of these purposes may be adequately served with task inventories at a general level of specificity, while others demand greater specificity. Yet, United States Air Force experience suggests that more detailed task lists are most productive.

EVOLUTION OF TASK INVENTORIES IN THE USAF

Over the past 11 years of operational occupational analysis in the US Air Force, more than 300 occupational areas have been analyzed and described. This effort required the writing of over 150,000 tasks which were administered to over 700,000 job incumbents.

In the early years of the operational experience, task inventories followed the model established through ten years of research of occupational analysis techniques. Since a major objective, if not the primary one, of early task lists developed during the research period was to support the Air Force classification process, the early task lists tended to be less detailed. The average number of tasks was about 350, and rarely did a list exceed 500 tasks. The broadly-written tasks contained in these inventories provided information about the subdivisions of the specialty upon which classification decisions could be made.

Other users of the data soon developed, primarily training managers and curriculum development personnel. These users began to request more detail. Presently task lists for the simpler specialties range from 350 to 600 tasks; for the more complex specialties they may average as many as 1000-1200 tasks. The longest USAF inventory, used to describe the variety of jobs in the Communication-Electronics Officer Utilization Field, contained 1,435 tasks. So far, there is no evidence that these longer inventories, if carefully constructed, have any deleterious effects on the stability of incumbent responses.

These longer, more detailed inventories provide more complete information for training decisions and at the same time provide specific information for later users of the data. These most recent users include: promotion testing; management engineering; maintenance engineering; and personnel research into such areas as aptitude requirements, job satisfaction, and job difficulty.

Since the goal of these longer inventories is more precise data, it is essential to apply the four fundamental criteria for developing task inventories.

FUNDAMENTALS OF TASK DEVELOPMENT

In 1967 Air Force Human Resources Laboratory Technical Report PRL-TR-67-11, Morsh and Archer published a Procedural Guide for Conducting Occupational Surveys in the United States Air Force. This guide remains the single best source of task writing procedures, and the criteria described below are readily found in that source. This paper intends to elaborate the criteria in light of the 150,000 plus tasks that have been written during the intervening years.

In developing tasks to describe an occupational field, the occupational analyst is charged with writing tasks that meet the following fundamental criteria:

a. Each task is time-ratable -- that the job incumbent can reasonably estimate the relative amount of time he or she spends on each task. This criterion normally eliminates tasks that begin with such words as "insure, have responsibility for" and "understand", which make it difficult or impossible to determine the relative time devoted to this activity.

b. Each task communicates in the language of the specialty. The task statement must be clear so that it is easily understood by career field incumbents, the people who must answer the questionaire. Terminology consistent with current usage in the career field leaves less chance for error or differing interpretations of task statements.

c. Each task is mutually exclusive of other tasks in the inventory; that is, whether a job incumbent indicates that he or she performs a task must be independent of his or her performance of all other tasks in the inventory.

d. Each task will differentiate among workers where actual task performance differs because of such factors as differences of jobs, experience level (apprentice, journeyman, technician), organizational level (command, staff, base, flightline, or shop), and whether or not the person is a supervisor.

DISCUSSION

In this section we will elaborate the four fundamental criteria for task development with examples, then examine the level of detail and its influence on the last two criteria. Finally, we will look at questions which serve to guide the task developer in setting the level of detail in a particular occupational survey.

231

Time-Ratable

Tasks must be time-ratable in order to depict clearly the relative time spent on a particular task. Words or phrases which are not time-ratable often creep into inventories if this basic criterion is not stressed. Examples of words and phrases which may not be time-ratable include the following: "insure, assure, assist, control, monitor, coordinate, recommend, determine, know how to, understand, have knowledge of," and "have responsibility for." These words and phrases are vauge and can prevent the respondent and the occupational analyst from determining when the task started and finished. In addition, some of these examples are not behaviors, but rather knowledges, like the word "understand." Some of these same words, however, can be used as time-ratable tasks depending on the career field; for example, while "to monitor supply accounts" may not be a time-ratable task for a supply apprentice, "monitoring the scope" may well be a time-ratable task for a weapons controller who actually does monitor a radar scope for six hours of his or her eight-hour day. Also, the word, "assist", can refer to a specific task in medical specialties such as "assist the surgeon" in operating room procedures, while in other career fields the word "assist" is very vague. The same word, "assist", in the machinist shop could mean stand close by and watch, set up the equipment, hold the equipment in place, clean up afterward, or actually do the task under supervision.

Communicates in Language of Specialty

Each task must communicate in the language of the specialty. To reduce the possibility for error or misinterpretation, USAF experience indicates that it is clearest to construct inventories using the current language of the career field. Terms used in the daily work of the career field have a definite meaning to incumbents. In addition, there are certain dangers in depending on an external source of definition, like a glossary of verbs. For example, if a glossary is used that depends on some impersonal source of vocabulary other than the definitions in common usage by career field members, respondents could make the following errors which would lead to unreliable task ratings: forget to read or simply skip reading the glossary, forget the glossary definition or confuse the provided definition with the common usage version. We have found that more solid and stable responses are obtained by using the language respondents use every day on their job.

It takes skilled occupational analysts to differentiate subtle shades of meaning which exist in career field vocabulary and clarify task statements in such a way as to prevent misinterpretation. For example, in some career fields the word "troubleshoot" means to isolate the problem, whereas in other career fields the same word means to both find and fix the problem. In this example, if a standard glossary definition differed from usage, the word, and consequently the task would be subject to much misinterpretation, in unreliability of the resulting responses.

Mutually Exclusive of Other Inventory Tasks

Each task must also be mutually exclusive of other tasks in the inventory. If two or more tasks are mutually dependent (that is, if one task is performed, the other task must also be performed, or vice versa) these tasks would be more correctly called subtasks or task elements of a parent task. For example, in a weather observer inventory the tasks, "determine wind speed" and "determine wind direction", are really subtasks of the parent task, "make wind observations", since every time the observer does one task he or she must do the other task. The parent task, then, could more succinctly describe both activities, and therefore shorten the inventory. Another reason for dropping the subtasks from the inventory is that mutually dependent tasks may falsely inflate the relative time spent in a duty area by forcing the respondents to indicate multiple responses for essentially the same task. Also, if a parent task is used, no information regarding percent members performing would be lost by dropping the subtasks in favor of the parent task.

In CODAP programs, each task is valued equally, even if it is infact a subtask and not mutually exclusive of other tasks in the inventory. When two subtasks are used instead of their parent task, responses of time spent and percent members performing are recorded on two tasks instead of one. If two subtasks are used instead of the one parent task, groups performing the parent task could appear more similar in the cluster-merger diagram. Individuals or groups responding that they do not perform the parent task could appear more different than they would have had they marked only one task negatively, rather than the two subtasks. Thus, a more representative picture of the career ladder's structure can be obtained by using a parent task, rather than its subtasks.

Differentiates Among Career Field Members

Each task must differentiate among career field members where actual task performance differs. For example, if an apprentice can do only parts of a task under supervision, the journeyman can do the entire task, and the technician can supervise the task as well as perform it, the task inventory should include items which enable the occupational analyst examining the cluster-merger diagram to make these distinctions. In order to distinguish between groups which make up an occupational specialty, the tasks must be written at a sufficiently specific level of detail. For example, if an inventory only lists journeyman tasks, excluding supervisory and apprentice tasks, most members of the career field will not be separated by their performance level. If the inventory does not allow respondents to choose tasks which distinguish the levels, the occupational analyst scrutinizing the cluster-merger diagram cannot find these differences.

Level of Detail

In terms of level of detail, the last two criteria are the most important. As we have seen above, if two (or more) tasks are mutually dependent, or if they do not differentiate the level of work, what will be learned from data collection on these items is likely to be spurious. Differentiation is critical for defining the various kinds of jobs which make up a specialty. The key to differentiation resides in the level of specificity of tasks. Also, if the tasks are mutually exclusive, task differentiation is enhanced. In describing tasks, it is essential to first determine if any activity consists of concomitant elements in which the activity is a parent task. This characteristic, alone, assures differentiation. But just assuring differentiation in today's environment, when longer inventories are needed, is not sufficient.

In many cases it may be necessary to find some way to combine parent tasks, which would normally stand alone in an inventory. For example, the combination of parent tasks may be necessary because an inventory is too long. In this case, an additional criterion for task writing is essential: the occupational analyst should determine if any tasks, which are being considered for combination are performed differently by job incumbents. That is, whether they always exist concomitantly at any given job location, job experience level, etc. If parent tasks do exist concomitantly at all levels and locations, then the level of detail may be set at a more general level and the parent tasks may then be combined. In these longer inventories where space is at a premium, tasks which have high similarity, or tasks which could be accomplished without additional training, can be combined into a more general and inclusive task. For example, instead of listing all 150 preflight inspection checklist items separately, the task, "conduct preflight inspection on (Type of Aircraft)" could be used if the preflight inspection is conducted the same way at all locations and experience levels but differs by aircraft. In this case, differentiation between those parent tasks is not necessary, even though the tasks may not be mutually dependent.

This new criterion, or exception to the rules of mutual exclusivity and differentiation, then may help shorten today's longer multiladder inventories. The level of detail can be adjusted according to this criterion without causing spurious data collection which results from failing to follow the four fundamental criteria.

SUMMARY

As occupational analysis becomes more sophisticated, the length of occupational survey task inventories have become longer The added length results from impetus to meet the following objectives: to describe tasks at the lowest level of work activities which describes a complete and inseparable operation, to provide technical training schools with the most useful data to structure their courses, and to best describe career field structure to classification interests by multi- or cross-ladder surveys. Longer surveys make critical four fundamental criteria for describing occupational survey tasks. These criteria are (1) each task must be time-ratable, (2) each task must communicate in the language of the specialty, (3) each task must be mutually exclusive of other tasks in the inventory, and (4) each task must differentiate among workers where actual task performance differs. Compromise between these criteria is often necessary in the practical world. The appropriate level of detail is determined by carefully ballancing criteria three and four. Setting the level of detail at the appropriate point maximizes the information to be gained from task inventories and minimizes the length to provide accurate data to users of the occupational survey program.

REFERENCES

Ammerman, Harry L. Performance Content for Job Training, VOL 2, "Stating the Tasks" Columbus, Ohio: Ohio State University, The Center for Vocational Education, March 1977.

Melching, W. H. & Borcher, S.D. Procedures for Constructing and Using Task Inventories (R&D Series No. 19) Columbus: Ohio State University The Center for Vocational Education, March 1973.

Morsh, J. E. & Archer, W. B. "Procedural Guide for Conducting Occupational Surveys in the United States Air Force" (PRL-TR-67-11) Lackland Air Force Base, TX: Aerospace Medical Division (ADSC), Personnel Research Laboratory, September 1967 (MTIS No. AD-664 036).

TWO APPLICATIONS OF OCCUPATIONAL SURVEY DATA IN MAKING TRAINING DECISIONS

Capt. David S. Vaughan
ATC Technology Applications Center

In this era of ever-tightening budgets, it has become extremely important that formal training content match, as closely as possible, actual job requirements. We can afford neither overtraining, which wastes training resources, nor undertraining, which increases the on-the-job training load and detracts from primary mission accomplishment. One very useful source of information for constructing job-relevant training programs in the Air Force is the occupational survey. Occupational surveys are accomplished on a routine basis for most Air Force enlisted job specialties by the USAF Occupational Measurement Center. Procedures used in these occupational surveys are described in Morsh and Archer (1967).

Data available to trainers from occupational surveys include the percent of airmen in a specialty (or in an identifiable subgroup of the specialty) who perform any given task, the relative time spent on each task, and task learning difficulty. As may be seen, this sort of information can be very useful for making training decisions. However, several important questions are not answered. In some job specialties, the criticality of a task plays an important role in determining training requirements. Task criticality is not directly assessed in conventional occupational surveys and may not have a close relationship with percent members performing or the other normal occupational survey variables. Consider, for example, the 571X0, Fire Protection, Air Force job specialty. In this specialty the most critical tasks and those for which training is most needed, such as putting out fires, are tasks that, hopefully, are seldom actually performed on the job. A second major question concerns the procedures which should be used to combine data on the several conventional occupational survey variables into one index for ranking tasks for training. For example, if task A has high percent members performing and moderate difficulty, while task B has moderate percent members performing and high difficulty, which should receive more emphasis in training? Without guidance concerning how to combine the occupational survey variables in making training decisions, attention may be focused on one of the variables to the exclusion of the others, or an arbitrary combination rule may be used which is less than optimal.

Recently, the Air Force Human Resource Laboratory (AFHRL) completed a research project which was designed to provide solutions to the two problems outlined above. Progress reports concerning this research have been presented at several recent conferences (Mial & Christal, 1974; Mead, 1975; Stacy, Thompson, & Thomson, 1977). Mr. Hendrick Ruck (Ruck, Thompson, & Thomson, 1978) will present a detailed description of this research and its results at this year's Military Testing Association convention. The present report is concerned with the application of AFHRL's recently-completed training research, along with conventional occupational survey data, in an actual Air Force training environment. Therefore, a detailed description of those research results will not be given here.

My organization, the Air Training Command (ATC) Technology Applications Center, is concerned with the application of research within ATC training. We apply research on a test basis, evaluate the success of the application, and if appropriate, assist in the full scale implementation of the research. With the AFHRL training research results, we are conducting two separate applications projects. In the first project, occupational survey data are being used for the narrow purpose of revising an existing apprentice-level resident training course. By contrast, the second project is concerned with determining total training requirements--both resident and on-the-job training (OJT)--for an entire Air Force job specialty. The specific procedures being field-tested in both of these projects will be discussed below. First, the AFHRL research results which are being field-tested will be summarized.

The main product of the AFHRL research being discussed here is a new occupational survey scale--field recommended training emphasis. This scale is illustrated in figure 1. Data on the scale are gathered by having senior noncommissioned officers in the job specialty under consideration rate each task. The ratings indicate the extent to which emphasis should be placed on each tas in formal training for first assignment--apprentice level--airmen. The ratings do not, however, distinguish among various forms of formal training, such as resident, field training detachment, or OJT. Two main reasons exist for this lack of distinction. First, on a logical basis, NCOs in the field do not have certain types of information--resource availability at technical training centers, for example--which are important in such decisions. Secondly, Mead (1975) showed that field NCOs have low agreement concerning resident training vs OJT. In contrast, the AFHRL research showed that, in most job specialties, data gathered on the field recommended training emphasis scale have high interrater agreement. Furthermore, the AFHRL research showed that ratings on this scale were predictable with great accuracy from task data on the various factors which the educational literature tells us should be important in making training decisions. In summary, the research showed that, in most job specialties, this new field recommended training emphasis scale is both reliable and valid.

Course Revision Project

The first of the two applications projects which we are conducting with this new field recommended training emphasis data has the relatively narrow goal of revising (or constructing an apprentice-level resident training course. The procedures being tested in this project assume that certain important decisions have already been made, at least tentatively. The first assumption is that an apprentice-level resident training course of some sort will exist in the particular job specialty being examined. Secondly, it is assumed that the "audience" for this course has already been defined. The previously-defined "audience" might, for example, be all airmen entering the job specialty or airmen entering the specialty who will be assigned to the Strategic Air Command. Finally, it is assumed that some general decisions have been made concerning jobs to be performed by airmen at various skill levels in the specialty (in particular, that an acceptable Specialty Training Standard STS exists). Of course, it is recognized that the outcome of the procedures being tested in this project may result in modification of these decisions.

The goal of these procedures is to translate, in a simple and direct manner, occupational survey data into course content. The procedures should allow each topic covered in the course to be traced back to one or more tasks which were identified for inclusion.

The first step of the procedures is to select tasks for inclusion in the course. For this purpose, a special occupational survey printout is provided. This printout is illustrated in figure 2. On the printout, all tasks are listed in order of their field recommended training emphasis. Beside each task is printed the field recommended training emphasis, the percent of first-assignment airmen performing the task, and the task difficulty. Using this printout, training personnel consider each task for inclusion in the course. In general, tasks with high field recommended training emphasis are the ones which should be included in the course. However, it is recognized that many other considerations may also be important in determining course inclusion for any given task. For example, a task with high field recommended training emphasis might be excluded from resident training if that task has low difficulty and/or percent members performing, or if equipment necessary to train that task cannot be made available at the training center. Consideration might be given to including a task with low field recommended field training emphasis if that task has high difficulty and percent members performing or if that task has a great deal of content overlap with other tasks already included in the course. When available, information from sources other than occupational surveys should also be used in making task decisions. For each task, the reasons for the decision to include or exclude are documented in a brief note beside the task on the printout. Certain rules of thumb are provided to simplify this documentation requirement. First, any task whose field recommended training emphasis is at least one standard deviation above the mean requires written documentation only if that task is to be excluded from the course (i.e., high field recommended training emphasis is sufficient to include a task unless other considerations dictate that the task be excluded). Any task whose field recommended training emphasis is at least one standard deviation below the mean requires written documentation only if it is to be included in the course (i.e., low training emphasis is sufficient for exclusion, unless other considerations are important). Only tasks whose field recommended training emphasis is within one standard deviation of the mean require documentation for either inclusion or exclusion. Training emphasis data does not provide an unambiguous answer for these "middle range" tasks, and other considerations will always be important.

The next step involves task analysis--determination of the skills and knowledges required to perform the task and, thus, behavioral objectives for the course. Student evaluation instruments--course examinations--are also constructed in this step. The original intent was for the task analysis to be accomplished through conventional Air Training Command procedures. However, at the same time that we were field-testing the procedures described here, AFHRL personnel were field-testing a new handbook for task analysis. Details of the AFHRL task analysis procedure are described in a paper to be presented at this conference (Eschenbrenner, De Vries, and Ruck, 1978). The AFHRL experimental handbook was made available to course revision personnel to use as a part of our course revision project. The result of the task analysis will be a complete list of skills and knowledges required to perform each of the tasks identified for inclusion in the course. Behavioral objectives are also constructed as part of the task analysis. Although not part of the AFHRL task analysis procedure, course examinations will also be developed during this step. The result of this step will allow each skill and knowledge, behavioral objective, and test item or segment to be traceable back to one or more of the occupational survey tasks which were identified for inclusion in the course.

In the third step, course personnel identify groups of tasks with common subject matter--common knowledges and skills--in order to provide organization for the course. Then we will sum the difficulties for all tasks in each group and divide these sums by the sum of difficulties for all tasks being included in the course. The result of this procedure is an approximate measure of the relative amount of training time to be devoted to each group of tasks. The relative training time measure is a guide for course planning and may be overriden when appropriate. This procedure can, if desired, be carried out for each individual task. Finally, actual course matter will be developed using normal, currently followed procedures.

It is difficult to predict the effects of this procedure on curriculum content. The course may be lengthened, shortened, changed in other ways, or remain the same. In any case, the reason for using occupational survey data in curriculum development is to insure that courses are closely aligned with the survey results and with actual job requirements. Even if a course is not changed appreciably, indicating that it was previously aligned with survey results, systematic use of survey data in revision of the course is desirable because it will provide evidence of this alignment. Therefore, the first criterion for success of the proposed procedure is that its application should result in courses that are no less closely aligned with occupational survey data than conventionally developed courses. This criterion will be evaluated by examining the relationship of occupational survey data to course content before and after revision. In addition, test items developed under the proposed procedure will be administered to graduates of the old and revised courses for follow-on comparisons. Significant revisions in course content should be reflected in the knowledge levels of course graduates. If graduates of the old course perform as well as graduates of the revised course on these new items, the occupational survey data probably had no major impact on the course content. Conversely, large test score difference could indicate significant revisions in course content. It is recognized that factors other than use of occupational survey data can result in score differences between graduates of old and revised courses. However, the test score data will be interpreted in light of other data to be gathered. For example, if large test score differences are found, but the POIs for the old and revised courses are identical, the test score differences are probably not due to use of occupational survey data in course development.

Two other criteria which should be met for the proposed procedure to be considered successful are efficiency and acceptability to users. To be considered efficient, any additional resources required by the proposed procedure should, in the judgment of users, be counterbalanced by benefits over and above those obtained by using conventional procedures. To measure possible additional resources required by the new procedure or resource savings obtained under the new procedure, users will rate the time and other resources required under the new procedure relative to that required under conventional procedures. Users will also rate the benefits of the new procedure relative to those conventional procedures, as well as the relative acceptability of the new procedure. This resource and benefit data, as well as data gathered to meet the first criterion, will be made available to the users, who will rate whether any additional benefits obtained under the new procedure are worth possible additional resource requirements. In order for the application to be considered successful, the new procedure must be at least as efficient and as acceptable as currently-used procedures.

We are currently testing these procedures in two job specialties--19333, Apprentice Radio Operator; and 91130, Apprentice Aerospace Physiology Specialist. In both of these courses, the task selection process was completed in less than one day. Course personnel are currently working on the second step--the task analysis. We had hoped to be able to report here some results from our formal evaluation of these procedures. However, the revision efforts are not far enough along in either of the two courses for much data to be available. We can say that course personnel found the task selection process to be reasonably efficient and are finding the results to be useful. The only real problem encountered so far concerns the training emphasis cutoff value below which written documentation is required only for a task to be excluded from the course. Field recommended training emphasis values have an extreme positive skew. Out of an inventory with over 500 tasks, only 50 or 100 may have very high recommended training emphasis values.

Therefore, the mean training emphasis is low, and one standard deviation below the mean is extremely low. Based on our experience to date, it appears that this cutoff should be set higher than one standard deviation below the mean, perhaps at an absolute value of 2.0.

Construction of Specialty Training Standards

An Air Force Specialty Training Standard (STS) defines training requirements for an entire career ladder. First, an STS serves as a specification document for formal training. Second, it is the basis for preparation of Career Development Courses (CDCs). Third, an STS is a guide for local OJT programs and for preparation of Job Proficiency Guides used in OJT. An STS contains information concerning the topics for which training is to be provided at each skill level (apprentice, specialist, and technician) in an Air Force Job Specialty (AFS). In addition, information is provided concerning the degree of training to be provided in OJT and in formal training courses and concerning reference material which may be used in training. A well-constructed STS, therefore, provides a comprehensive description of training requirements in an entire AFS, including data concerning what tasks are to be trained at each skill level and the extent of the training to be provided. A sample page from an STS is illustrated in figure 3. Each subject area on an STS has skill-knowledge codes which indicate the amount of training to be provided at each skill level. Figure 4 contains a key for these skill-knowledge codes.

The purpose of this application project is to develop and field test a systematic procedure for applying occupational survey data in constructing STSs. Algorithms will be tested for selection of tasks to appear on an STS, for identification of tasks for resident training and for assignment of skill-knowledge codes. The new field recommended training emphasis factor will be used in addition to normal occupational survey data. This project is a direct follow-on to that of Ruck, Dineen, and Cunningham (1977).

Under this STS construction procedure, a number of decisions are made using arbitrary, although reasonable, cutoff values. It is recognized that these cutoffs may not be appropriate in certain circumstances, and that information not contained in the occupational survey data may also be relevant in making STS decisions. Therefore, a manual override option is allowed at each decision point concerning task selection and skill-knowledge coding. However, reasons for manual override will be documented, and approval will be obtained from appropriate authorities.

The first step involves selection of the tasks to appear on the STS. Also, in this step, tasks will be matched with particular skill levels in the AFS. STS task statements will be taken directly from the occupational survey task list. The use of occupational survey task statements has several advantages. First, a great deal of labor can be saved, since a well written task inventory provides, ready-made, a detailed, behaviorally-oriented breakdown of all job activities in a particular career ladder. Secondly, use of occupational survey tasks on an STS eliminates the problem of relating occupational survey data, based on one job breakdown, to an STS, which is usually based on a different breakdown. Finally, use of occupational survey task statements on an STS eases the establishment of the relationship between occupational survey data and other training documents, such as course Plans of Instruction, which are based on STSs. The rule followed for selection of tasks is that any task with at least 10 percent members performing at any skill level will appear on the STS and will be matched with that skill level. In addition, tasks will be matched with all levels above such a skill level (for example, the five skill level will include all tasks matched with the three skill level, as well as tasks not performed at the three skill level).

The second step involves selection of tasks for inclusion in an ABR course. The field recommended training emphasis task factor will be used to make these decisions. Any tasks whose field recommended training emphasis exceeds a minimum cutoff value on the training emphasis task factor will be that which is equivalent, in the regression sense, to 30 percent members performing among first assignment airmen.

The last step involves assignment of skill-knowledge codes to the STS. Two different procedures will be tested for skill-knowledge code assignment. A separate STS will be constructed using each of the two procedures, allowing a direct comparison of the procedures. Examples of STSs constructed under these two procedures are contained in figures 5-6.

(1) Under the first procedure, a skill-knowledge code will be assigned to each task for each skill level with which the task is associated. These skill-knowledge codes will be assigned through currently followed procedures.

(2) The second skill-knowledge code assignment procedure is based on the "go-no go" philosophy. Under this philosophy, an OJT trainer signs off an STS area when an airman reaches the "go" level of performance in that area. No gradations of performance or knowledge are recognized in OJT beyond the "go" and "no go" levels. Herein, assignment of skill-knowledge codes to STS areas is assumed unnecessary for OJT purposes. However, skill-knowledge codes will still be needed to reflect partial training on particular tasks which may be given in an ABR course. Therefore, the following skill-knowledge code assignment procedure will be used only to assign codes for tasks included in an ABR course: Any task whose field recommended training emphasis exceeds a value which is equivalent, in the regression sense, to 50 percent members performing will be assigned the 2b code. All other tasks included in the ABR course will be assigned the 1a code. As with all other automated decisions, those made in this step may be modified through manual override.

Under the conventional skill-knowledge approach, the difference among the various skill levels are mainly a matter of degree--of how well airmen can perform various tasks. Under the "go-no go" approach to be tested, differences among skill levels are not a matter of "how well", but a matter of what tasks are performed. For example, a specialist-level airman can perform all tasks that an apprentice-level airman can perform, as well as some additional tasks.

An evaluation will be conducted of these experimental STS construction procedures. The evaluation will involve comparisons among three STSs: a "go-no go" experimental STS, an experimental STS with conventional skill-knowledge coding, and the conventional STS. The most important criterion in the evaluation will be acceptability to users. This criterion will be measured through surveys of all classes of users--formal trainers, OJT trainers and supervisors, career field functional managers, etc. Another criterion will be the ease with which the proposed procedures can be followed. This will be evaluated through surveys of personnel who are directly involved with the construction of the experimental STSs.

These STS construction procedures are currently being applied in three job specialties. We had hoped to be able to present some results from the formal evaluations, but STS construction is not yet complete in any of the specialties. However, in one of the specialties, 911X0, Aerospace Physiology, one of the experimental STSs is complete (with conventional skill-knowledge coding), and the other is nearing completion. We are able to offer some observations from that experience. The actual selection of tasks was completed in less than a day by training personnel. The remainder of the procedure (assignment of skill-knowledge codes, grouping tasks into subject areas, revising some tasks, etc.) was completed with about 15 hours of labor. Course personnel are very satisfied with the results. Also, an Instructional Systems Design specialist who serves as a consultant to the entire school in which the Aerospace Physiology courses are located, reviewed the experimental STS and indicated that he liked the STS. It seems fair to say, based on this preliminary information, that formal training personnel like the experimental STSs, at least the conventionally skill-knowledge coded STS. However, no information is yet available concerning how other users, such as OJT trainers will like the experimental STS.

Summary

Two sets of procedures for using occupational survey data, including data on the new field recommended training emphasis scale, in training decision-making are currently being field-tested. Both sets are designed to provide simple and efficient methods for translating occupational survey data into training content. The first set of procedures have the limited goal of revising (or constructing) an apprentice-level resident training course. The second set has the more ambitious goal of determining all training requirements, both resident and OJT, in an entire career ladder. Results of the planned formal evaluations are not yet available. However, both sets of procedures are currently being applied in several job specialties; the experience obtained to date suggests that both sets of procedures will be useful.

REFERENCES

Eschenbrenner, A. J., De Vries, P. B., & Ruck, H. W. Methods for Collecting and Analyzing Task Analysis Data. Paper presented at the 20th annual conference for the Military Testing Association, U. S. Coast Guard, Oklahoma City, OK, October, 1978.

Mead, D. F. Determining Training Priorities for Job Tasks. Paper presented at the 17th annual conference for the Military Testing Association, U. S. Army, Indianapolis, IN, September, 1975.

Mial, R. A. & Christal, R. E. A Determination of Training Priorities for Vocational Tasks. Paper presented at the Psychology in the Air Force symposium, U. S. A. F. Academy, April, 1974.

Morsh, J. W. & Archer, W. B. Procedural Guide for Conducting Occupational Surveys in the United States Air Force (PRL-TR-67-11).

Ruck, H. W., Dineen, R. T., & Cunningham, C. C. Applying Occupational Survey Data in Instructional Systems Development. Paper presented at the 19th annual conference for the Military Testing Association, U. S. Air Force, San Antonio, TX, October, 1977.

Ruck, H. W., Thompson, N. A., & Thomson, D. C. The Collection and Prediction of Training Emphasis Ratings for Curriculum Development. Paper presented at the 20th annual conference for the Military Testing Association, U. S. Coast Guard, Oklahoma City, OK, October, 1978.

Stacy, W. J., Thompson, N. A., & Thomson, D. C. Occupational Task Factors for Instructional Systems Development. Paper presented at the 19th annual conference for the Military Testing Association, U. S. Air Force, San Antonio, TX, October, 1977.

FIELD RECOMMENDED TRAINING EMPHASIS SCALE

CHECK EACH TASK FOR WHICH YOU RECOMMEND FORMAL TRAINING FOR FIRST-TERM AIRMEN.

RATE ONLY THE TASKS YOU CHECKED TO INDICATE HOW MUCH FORMAL TRAINING EMPHASIS YOU RECOMMEND FOR FIRST-TERM AIRMEN.

1 EXTREMELY LITTLE
2
-
-
-
5 AVERAGE
-
-
-
-
8
9 EXTREMELY HEAVY

COURSE REVISION PROJECT
COMPUTER PRINTOUT

		SEQUENCE NUMBER	EMPHASIS	DIFFICULT	% MEMBERS PERFORMING
F 103	SERVE AS INSIDE OBSERVER ON TRAINING CHAMBER FLIGHTS	1	7.99	3.89	96.9
F 104	SERVE AS LOCK OPERATOR ON TRAINING CHAMBER FLIGHTS	2	7.79	4.54	93.8
F 101	SERVE AS CHAMBER OPERATOR ON TRAINING CHAMBER FLIGHTS	3	7.74	3.97	93.8
	⋮		⋮		
H 193	PERFORM STRUCTURE TESTS OF PRESSURE SUIT GLOVES	144	2.62	4.77	3.1
H 236	SUIT UP CREW MEMBERS WITH PRESSURE SUITS	145	2.59	4.39	4.6
E 87	PROOFREAD CORRESPONDENCE, REPORTS OR FORMS	146	2.51	4.01	23.1
	⋮		⋮		
A 2	ACT AS TRAINING PROGRAM ADVISOR AT STAFF LEVEL	342	.10	5.89	10.8
D 66	DEVELOP RESIDENT COURSE CURRICULA MATERIALS	343	.10	5.59	3.1
B 23	CONDUCT STAFF MEETINGS	344	.06	4.91	3.1

1.	PROFICIENCY LEVEL, PROGRESS RECORD AND CERTIFICATION								
	2. 3 SKILL LEVEL			3. 5 SKILL LEVEL			4. 7 SKILL LEVEL		
TASKS, KNOWLEDGE AND STUDY REFERENCES	A AFSC /Crs	B Date OJT Started	C Date Compld & Trainee's Supervisor's Initials	A AFSC	B Date OJT Started	C Date Compld & Trainee's Supervisor's Initials	A AFSC /Crs	B Date OJT Started	C Date Compld & Trainee's Supervisor's Initials
20. ANIMAL SERVICE AND ZOONOTIC DISEASE CONTROL									
SR: AFMs 125-5 (chap 4, vol 1), 160-12, 160-37, 160-48; AFRs 125-9, 161-4, 163-4, 163-11, 168-10; AFP 163-1-3, 163-10 (sec A & B); Catcott, E. J. Animal Hospital Technology, American Veterinary Publications (AVP), 1971; Benbrook, E. A., and M. W. Sloss, Veterinary Clinical Parasitology, Iowa State University Press, 3rd ed, 1961									
a. Principles of animal care, management, medicine and surgery	A			B			C		
b. Principles of identification and control of zoonotic and other diseases of animals (including controlling entry of foreign animal diseases into the US)	A			B			C		
c. Assist in the zoonoses control program	1a/a			3c			4c		
d. Assist in the management, veterinary care, treatment and necropsy of government owned animals	1a/a			3c			4c		
e. Prepare reports and maintain records pertaining to veterinary care of:									
(1) Privately owned animals	1a			3b			4c		
(2) Government owned animals	1a			3b			4c		
f. Perform laboratory and/or clinical procedures related to control of animal and zoonotic diseases	2b			3c			4c		
g. Procedures for evaluation and decontamination of military working dogs exposed to nuclear, biological or chemical agents	-			C			C	NO ADVANCED COURSE	
21. ANIMAL TECHNICIAN SPECIALTY (For personnel assigned duties as an Animal Technician, SEI 491, exclude from consideration in development of SKT and CDC)									
a. Occupational health and safety									
SR: AFP 161-25; AFRs 92-1, 127-101, 127-4, 127-6, 127-12, 160-56, (chap 7) 160-57, 160-132 161-6, 161-8, 161-18, 161-24									
(1) Injury and zoonotic disease hazards in the research animal colony	-			C			D		
(2) Apply appropriate occupational safety practices	-			3b			4c		
b. Medical terminology									
SR: AFR 160-56 (chap 2); American Association for Laboratory Animal Science Pub 67-3, Manual for Laboratory Animal Technicians, 1967 (hereafter listed as AALAS Pub 67-3); Purina Manual; Worden, A. N. and Lane Petter, W. eds: The UFAW Handbook on the Care and Management of Lab Animals, 4th ed., The Universities Federation for Animal Welfare, 1972 (hereafter listed as the UFAW Handbook)									
(1) Medical terminology relating to anatomy and physiology	-			B			C		
(2) Disease	-			B			C		
(3) Surgery	-			B			C		
(4) Axenic animals	-			B			C		

PROFICIENCY CODE KEY

	SCALE VALUE	DEFINITION: The Individual
TASK PERFORMANCE LEVELS	1	Can do simple parts of the task. Needs to be told or shown how to do most of the task. (EXTREMELY LIMITED)
	2	Can do most parts of the task. Needs help only on hardest parts. May not meet local demands for speed or accuracy. (PARTIALLY PROFICIENT)
	3	Can do all parts of the task. Needs only a spot check of completed work. Meets minimum local demands for speed and accuracy. (COMPETENT)
	4	Can do the complete task quickly and accurately. Can tell or show others how to do the task. (HIGHLY PROFICIENT)
*TASK KNOWLEDGE LEVELS	a	Can name parts, tools, and simple facts about the task. (NOMENCLATURE)
	b	Can determine step by step procedures for doing the task. (PROCEDURES)
	c	Can explain why and when the task must be done and why each step is needed. (OPERATING PRINCIPLES)
	d	Can predict, identify, and resolve problems about the task. (COMPLETE THEORY)
**SUBJECT KNOWLEDGE LEVELS	A	Can identify basic facts and terms about the subject. (FACTS)
	B	Can explain relationship of basic facts and state general principles about the subject. (PRINCIPLES)
	C	Can analyze facts and principles and draw conclusions about the subject. (ANALYSIS)
	D	Can evaluate conditions and make proper decisions about the subject. (EVALUATION)

– EXPLANATIONS –

* A task knowledge scale value may be used alone or with a task performance scale value to define a level of knowledge for a specific task. (Examples: b and 1b)

** A subject knowledge scale value is used alone to define a level of knowledge for a subject not directly related to any specific task, or for a subject common to several tasks.

– This mark is used alone instead of a scale value to show that no proficiency training is provided in the course, or that no proficiency is required at this skill level.

X This mark is used alone in course columns to show that training is not given due to limitations in resources.

PREVIOUS EDITION IS OBSOLETE. REPLACES
ATC FORM 23A, JAN 75, WHICH IS OBSOLETE.

CTS/TENTATIVE STS PROFICIENCY CODE KEY

1.	2.	3.	
h. Teach post-flight chamber flight procedures	2b/1a	3c	3
i. Teach procedures during chamber flights	2b/1a	3c	3c
11. HYPOBARIC CHAMBER MAINTENANCE AND INSPECTION			
SR: T.Os. 43D8-3-1-101, 43D8-3-2-6			
a. Perform daily inspections of low pressure chambers	2b	3c	3c
b. Perform periodic inspections of low pressure chambers	2b	3c	3c
c. Perform special inspections of low pressure chambers	2b	3c	3c
d. Recharge batteries for emergency intercom systems	2b/-	3c	3c
e. Remove or replace flourescent tubes inside low pressure chambers	2b/-	3c	3c
f. Remove or replace operator panel instruments	2b/-	3c	3c
g. Remove or replace oxygen equipment items on low pressure chambers	2b/-	3c	3c
h. Add oil to vacuum pumps	2b/-	3c	3c
SR: T.Os. 34Y5-3-29-4; 34Y5-3-35-1			
i. Solder breaks in intercom wiring	2b/-	3c	3c
j. Prepare or maintain records on status or inspections of equipment	2b	3c	3c
SR: T.Os. 00-20-5, 00-20-7			
12. LIFE SUPPORT EQUIPMENT FUNCTIONS			
SR: AFP 160-5 (chap 13 and 14)			
a. Fit oxygen masks	3c/2b	3c	3c
SR: T.O. 15X-4-4-12			
b. Fit parachutes	2b/1a	2b	2b
SR: T.Os. 14D1-1-1, 14D1-2-1			

TASKS, KNOWLEDGES AND STUDY REFERENCES 1.	Lvl AFSC/C.. 2.	Lvl AFSC 3	Lvl AFSC/C.. 4
h. Teach post-flight chamber flight procedures	X/1a	X	X
i. Teach procedures during chamber flights	X/2b	X	X
11. HYPOBARIC CHAMBER MAINTENANCE AND INSPECTION			
SR: T.Os. 43D8-3-1-101, 43D8-3-2-6			
a. Perform daily inspections of low pressure chambers	X/2b	X	X
b. Perform periodic inspections of low pressure chambers	X/1a	X	X
c. Perform special inspections of low pressure chambers	X/-	X	X
d. Recharge batteries for emergency intercom systems	X/2b	X	X
e. Remove or replace flourescent tubes inside low pressure chambers		X	X
f. Remove or replace operator panel instruments		X	X
g. Remove or replace oxygen equipment items on low pressure chambers			X
h. Add oil to vacuum pumps			X
SR: T.Os. 34Y5-3-29-4; 34Y5-3-35-1			
i. Solder breaks in intercom wiring			X
j. Prepare or maintain records on status or inspections of equipment	X/2b	X	X
SR: T.Os. 00-20-5, 00-20-7			
12. LIFE SUPPORT EQUIPMENT FUNCTIONS			
SR: AFP 160-5 (chap 13 and 14)			
a. Fit oxygen masks			X
SR: T.O. 15X-4-4-12			
b. Fit parachutes	X/-	X	X
SR: T.Os. 14D1-1-1, 14D1-2-1			

THE STABILITY OVER TIME OF AIR FORCE ENLISTED CAREER LADDERS AS OBSERVED IN OCCUPATIONAL SURVEY REPORTS[1]

Walter E. Driskill, Ph.D.
and
Frederick B. Bower, Jr., Capt, USAF

USAF OCCUPATIONAL MEASUREMENT CENTER
OCCUPATIONAL SURVEY BRANCH
LACKLAND AFB, TEXAS 78236

A paper presented at the Military Testing Association Convention

30 October - 3 November 1978

---

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

253

THE STABILITY OVER TIME OF AIR FORCE ENLISTED CAREER LADDERS AS OBSERVED IN OCCUPATIONAL SURVEY REPORTS

Walter E. Driskill, Ph.D.
and
Frederick B. Bower, Jr., Capt, USAF

USAF Occupational Measurement Center
Occupational Survey Branch
Lackland AFB TX, 78236

A basic assumption behind the Air Force occupational survey has been that advances in technology and improvement in management procedures and techniques create over time, changes in the type of job performed within a given occupational specialty. Through the occupational survey, these changes could be identified and the appropriate updating of classification documents and training programs would then be made so that individuals in that occupation are trained and utilized in the most efficient manner. Research seems to indicate that the program has been pointed toward the identification of change in Air Force jobs since its early development days.

One objective of the Air Force program as described by Morsh (1964) is the identification of job changes and the determination of training needs. He determined this during reliability studies of the job inventory methods of occupational survey, although as Prien and Ronan (1971) point out, the logical research extensions are not reported. This was also the premise of Christal (1969) in his reliability studies of the job inventory. Both assumed that since reliability varies depending on the time interval between ratings, changes in the job survey would be noted over time. However, this early emphasis on identifying change in order to show the reliability of the survey instrument may have led those within the program away from identifying job stability. As pointed out by Driskill, Keeth, and Mitchell (1978), The USAF Occupational Measurement Center has now been in existence long enough to have resurveyed many enlisted career ladders for the second and sometimes third time. As such, our perceptions of how to approach the analysis of occupational survey data is changing. Like Morsh and Christal, we see the change in the areas of time requirements and task occurance, but we are also seeing stability in the job structure of many career ladders as evidenced in the recent surveys. Of the 76 occupational survey of enlisted career ladders surveyed between 1 January 1977 and 30 June 1978, 71 of the ladders were being resurveyed and 59 of these were found to have remained essentailly stable over the time since the previous survey. Seven career ladders were identified as having changed to some degree but none had changed to any great extent. No determination of stability or change could be made for the remaining five because either radical differences between formats of the survey instruments or different approaches to the job analysis by the survey analysts made comparisons too difficult. It should be pointed out that this comparison between surveys is now made as a routine part of every survey analysis. The determination of career ladder stability is made by the survey analyst based on the data collected. Nineteen analysts working independently of one another determined the stability of these 59 career ladders as a part of their normal job and not as any sort of special project or study.

To illustrate just how stable career ladders can appear, two such specialties will be used to display the various comparisons that can be made to determine stability over time between surveys. These career ladders were chosen for ease of data display and because the jobs performed in the specialties are readily understood both inside and outside the military community. The two career specialties chosen as examples are Dental Laboratory Personnel and Air Force Recruiters. Dental Laboratory Personnel are responsible for the fabrication and repair of dental prostheses such as complete dentures, partial dentures, bridges, and crowns. Air Force Recruiters are responsible for contacting, interviewing, and smoothly processing prospective applicants for active duty Air Force service.

The first comparison that can be made between surveys is that of career ladder structure. This is the job structure of the career specialty determined on the basis of what people are actually doing in the field. The job groups are determined throught computer analysis using the Comprehensive Occupational Data Analysis Programs (CODAP). The CODAP groups jobs according to similarity of respondents' responses to the job tasks performed and the amount of time spent performing those tasks. Table 1 depicts the comparison of the Dental Laboratory career ladder structure between the April 1974 survey and the Jun 1978 survey. Every job identified in the first survey can also be found in the career ladder structure in the current survey. The differences in groupings are merely a function of each survey analyst's preference in choice of reporting points. Some analysts prefer to report small individual job groups while others prefer to report larger job clusters.

Another point to be brought out on this slide is the decrease from the previous study in the number of personnel fabricating removable partial dentures and the increase in the current study of personnel fabricating crowns, bridges and porcelain products. As dental technology has improved the quality and appearance of prosthetic implants, demand for these products has increased while the use of removable partial dentures has decreased. However, some patients will always require removable partial dentures for one reason or another, so the job of their fabrication will not go away. Therefore, the job structure within the Dental Laboratory career ladder remains stable even though the number of personnel working in particular jobs has changed.

Table 2 shows the comparison of the career ladder structure between surveys of recruiter personnel. The job ladder here is remarkably similar considering the extensive revision and reorganization of the survey instrument used to collect the data for the current study. The improvements in the job inventory resulted in the identification of the Production Management and Classification Interviewer jobs. However, recruiter personnel revealed these jobs had existed at the time of the first survey, but tasks had not been included in that job inventory to capture them. This further tended to verify the stability of this specialty.

Another comparison made to determine career ladder stability is that of the percent time spent performing various duties of the job. Since none of the duty titles changed between development of the job inventories used to survey Dental Laboratory personnel, Table 3 provides a good example of this comparison. The differences in time spent fabricating and repairing removable partial dentures and fabricating procelain products was explained previously. The 1-24 month active federal military service group was chosen to further illustrate stability of the initial job assignment in this career ladder in that there is no carry-over of personnel from the previous survey to the current survey.

A comparison of the percent of members performing tasks between surveys is also used to determine stability of jobs over time. Table 4 shows this comparison of tasks for Dental Laboratory personnel with 1-24 months active federal military service. Again, despite the completely different makeup of each sample group, the percent of members performing each task is comparable.

Also shown on Table 4 is a comparison of the difficulty of each task between surveys. Task difficulty is determined by asking experienced personnel in the job specialty to rate each task in the survey instrument on the basis of how long it takes to learn to do the task. A nine point scale is used with "one" being a very small amount of time needed to learn the task to "nine" being a very large amount of time to learn the task. The ratings are then computer adjusted so that tasks of average difficulty have ratings of 5.00. Task difficulty ratings are accomplished for each survey and the sample chosen to perform the ratings is selected at random. Therefore, the high degree of similarity in task difficulty ratings is evidence that the perceptions of the difficulty of jobs within this particular career ladder have not altered over time.

Prior to 1974, task difficulty ratings were not adjusted. Rather the raw average scores were utilized. As Table 5 illustrates, even when comparing raw scores to adjusted scores, the order of task difficulty remains relatively the same.

The final comparison made for career ladder stability is that of job skill level. Table 6 depicts a different specialty than the previous examples but one chosen because it spans nearly 10 years between the first and the current surveys. As illustrated, 5-skill level Inventory Management Specialists have remained relatively constant in the percent of members performing the various tasks relative to their jobs. Only in the areas of operating data processing equipment has

256

there been a steady rise in the number of personnel performing those tasks. As the Air Force supply function became more automated such an occurance was naturally expected.

As shown, the determination that a career ladder is stable is more than just identifying like job groups. It is an in-depth comparison between surveys of not only the career ladder structure but a comparison by skill level and time in service groups plus task difficulty as well.

The implications of identifying so many stable career ladders are varied and complicated. Certainly classification and training personnel will be better able to manage their resources and training programs with this knowledge. However, these managers must not let themselves neglect stable career ladders. Even in the most stable of career areas, as technology improves and the Air Force acquires new and more sophisticated weapon systems and equipment, utilization patterns and training needs will change. Certainly stable career ladders need not be surveyed as frequently as they may have been in the past. However, we must remain responsive to changes in the field and always be prepared to provide timely data on any career ladder if the requirement arises. Certainly the verification of career ladder stability will allow survey analysts the time to broaden their horizons and explore the possibilities of other uses and applications of the survey data. However, analysts must never lose sight of the fact that the foundation of an occupation is the job structure, and that job structure has to be identified in order to properly interpret any of the other factors relating to the personnel performing in that career specialty. While the concept of career fields is utilized primarily by the military, as McCormick (1976) points out there are many civilian areas that could also be viewed as career fields. As such, job stability is very likely within the civilian community as well. Like military managers, civilian personnel utilizing occupational survey data must guard against identifying a stable job area and then failing to continue to monitor it for change in the future.

The apparent stability of the majority of jobs in the Air Force enlisted career structure has only recently been identified. There is much to do in this area before such data can be fully emploited. For example, job stability must be defined and objective criteria established so that stability may be determined. Even now though, the concept of stability within Air Force career ladders is impacting on the Occupational Survey Program and on the use of occupational data in classification training construction of career development courses, testing, and other USAF personnel programs.

AIR FORCE SPECIALTIES IDENTIFIED AS STABLE THROUGH OCCUPATIONAL SURVEYS CONDUCTED JANUARY 1977 THROUGH JUNE 1978

AIR FORCE SPECIALTY CODE	CAREER LADDER TITLE	CURRENT SURVEY	PREVIOUS SURVEY
114X0	AIRCRAFT LOADMASTER	JUN 77	SEP 71
242X0	DISASTER PREPAREDNESS	JUN 77	MAR 73
291X0	TELECOMMUNICATIONS OPERATIONS	FEB 77	OCT 70
303X3	AUTOMATIC TRACKING RADAR	NOV 77	JUN 73
305X4	ELECTRONIC COMPUTER SYSTEMS	DEC 77	DEC 72
306X2	TELECOMMUNICATIONS SYSTEMS/EQUIPMENT MAINTENANCE	FEB 78	JUN 73
316X0F	MISSILE SYSTEMS ANALYSIS (TITAN II)	NOV 77	OCT 73
316X1F	MISSILE SYSTEMS MAINTENANCE (TITAN II)	NOV 77	OCT 73
316X0G	MISSILE SYSTEMS ANALYSIS (WS-133AM/CDB)	JUN 78	OCT 73
321X2	WEAPON CONTROL SYSTEMS (F-5E)	DEC 77	JAN 73
321X2A	(F-106, ASQ-25 SYSTEMS)	DEC 77	JAN 73
321X2C	(F-106, ASQ-25 SUBSYSTEMS)	DEC 77	JAN 73
321X2N	(F-105D/F)	DEC 77	JAN 73
321X2P	(F-4C/D	DEC 77	JAN 73
321X2Q	(F-4E)	DEC 77	JAN 73
321X2S	(A-7D)	DEC 77	JAN 73
325X0	AUTOMATIC FLIGHT CONTROL SYSTEMS	OCT 77	MAR 72
328X4	AVIONIC INERTIAL AND RADAR NAVIGATION SYSTEMS	APR 78	APR 71
362X1	TELEPHONE SWITCH EQUIPMENT, ELECTRO/ MECHANICAL	MAR 78	FEB 72
423X2	AIRCREW EGRESS SYSTEMS	JAN 78	FEB 73
423X5	AEROSPACE GROUND EQUIPMENT REPAIRMAN	JUL 77	APR 71
431X1	TACTICAL AIRCRAFT MAINTENANCE		
431X1A	(A-7)	JUN 77	FEB 69
431X1C	(F/RF-4)	JUN 77	FEB 69
443X0G	MINUTEMAN MISSILE MECHANIC	JUL 77	SEP 71
472X0	BASE VEHICLE EQUIPMENT MECHANIC	JUN 78	JUN 72
472X1	SPECIAL PURPOSE VEHICLE MECHANIC		
472X1A	(FIRE TRUCKS)	JUN 78	JUN 72
472X1B	(REFUELING VEHICLE)	JUN 78	JUN 72
472X1C	(MATERIALS HANDLING EQUIPMENT)	JUN 78	JUN 72
472X1D	(TOWING AND SERVICING VEHICLES	JUN 78	JUN 72
472X2	GENERAL PURPOSE VEHICLE MECHANIC	JU[illegible] 78	JUN 72
473X3	VEHICLE BODY MECHANIC	JUN 78	JUN 72
511X0	COMPUTER OPERATIONS		
511X0A	(BURROUGHS SYSTEMS)	MAR 77	MAR 73
511X0B	(HONEYWELL SYSTEMS)	MAR 77	MAR 73
511X0C	(IBM SYSTEMS)	MAR 77	MAR 73
511X1	COMPUTER PROGRAMING		
511X1A	(BURROUGHS SYSTEMS)	MAR 77	MAR 73
511X1B	(HONEYWELL SYSTEMS)	MAR 77	MAR 73
511X1C	(IBM SYSTEMS)	MAR 77	MAR 73

AIR FORCE SPECIALTIES IDENTIFIED AS STABLE THROUGH OCCUPATIONAL SURVEYS CONDUCTED JANUARY 1977 THROUGH JUNE 1978

(CONTINUED)

AIR FORCE SPECIALTY CODE	CAREER LADDER TITLE	CURRENT SURVEY	PREVIOUS SURVEY
511X2	COMPUTER SYSTEM ANALYSIS AND DESIGN	MAR 77	MAR 73
542X0	ELECTRICIAN	OCT 77	MAR 73
542X0F	(TITAN II)	OCT 77	MAR 73
542X1	ELECTRICAL POWER LINE	MAY 77	JUN 72
542X2	ELECTRICAL POWER PRODUCTION	JUN 78	AUG 73
544X0	CRYOGENIC FLUIDS PRODUCTION	JUN 78	DEC 73
545X0	REFRIGERATION AND AIR CONDITIONING	SEP 77	MAR 71
547X0	HEATING SYSTEMS	SEP 77	MAR 71
554X0	REAL ESTATE-COST MANAGEMENT ANALYSIS	JUN 78	MAY 74
571X0	FIRE PROTECTION	APR 78	MAR 72
601X4	PACKAGING	MAY 78	OCT 73
602X0	PASSENGER AND HOUSEHOLD GOODS	MAY 78	OCT 73
602X1	FREIGHT TRAFFIC	MAY 78	OCT 73
622X1	DIET THERAPY	MAR 78	OCT 73
701X0	CHAPEL MANAGEMENT	MAY 78	DEC 73
901X0	AEROMEDICAL	MAR 77	NOV 71
982X0	DENTAL LABORATORY	JUN 78	APR 74
99500	RECRUITER	MAY 78	MAR 73

BIBLIOGRAPHY

Christal, R. E., Comments By The Chairman. In Proceedings of 19. Division of Military Psychology Symposium: Collecting. analyzing, and reporting information describing jobs and occupations. (77th Annual Convention of the American Psychological Association.) Lackland Air Force Base, Texas: Personnel Research Division, Air Force Human Resources Laboratory, September 1969, 71-72.

Driskill, W. E., Keeth, J. B., Mitchell, J. L., Differential Perception of Air Force Jobs. The Study of Air Force Jobs: Symposium Papers (6th Annual Psychology in the DOD Symposium, USAF Academy, Colorado, 22 April 1978). Lackland Air Force Base, Texas: USAF Occupational Measurement Center, 78-01, April 1978, 5-12.

McCormick, E. J., Job and Task Analysis in M. D. Dunnette (Ed), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally College Publishing Co., 1976.

Morsh, J. E., Job Analysis In The United States Air Force. *Personnel Psychology*, 1964, *17*, 7-17.

Prien E. P., and Ronan, W. W., Job Analysis: A Review of Research Findings. *Personnel Psychology*, 1971, *24*, 371-396.

TABLE 1

COMPARISON OF CAREER LADDER STRUCTURE BETWEEN SURVEYS OF AFS 982X0 DENTAL LABORATORY PERSONNEL

APR 74 SURVEY (N=501)	JUN 78 SURVEY (N=532)
COMPLETE DENTURE CLUSTER (N=191)	BASE DENTAL LAB PERSONNEL (N=307)
WORKING SUPERVISION CLUSTER (N=103)	
ORTHODONTIC JOB TYPE (N=8)	ORTHODONTIC APPLIANCE SPECIALISTS (N=9)
CROWN AND BRIDGE CLUSTER (N=56)	CROWN AND BRIDGE FABRICATORS (N=97)
	PORCELAIN FABRICATORS (N=15)
METAL FINISHING CLUSTER (N=51)	REMOVABLE PARTIAL DENTURES FABRICATORS (N=44)
WAX UP CLUSTER (N=22)	
AREA LAB SUPERVISION CLUSTER (N=33)	DENTAL LAB MANAGERS (N=29)

TABLE 2

COMPARISON OF CAREER LADDER STRUCTURE BETWEEN SURVEYS OF RECRUITER PERSONNEL

MAR 73 SURVEY (N=1665)	MAY 78 SURVEY (N=1615)
GENERAL RECRUITER CLUSTER (N=1192) [illegible] RECRUITER CLUSTER (N=18) [illegible] RECRUITER (N=7)	RECRUITER SALESMEN (N=1127)
SUPERVISORY CLUSTER (N=190)	RECRUITER MANAGEMENT PERSONNEL (N=166)
AF LIAISON NCO (N=61)	AFEES LIAISON NCOs (N=166)
ADVERTISING AND PUBLICITY CLUSTER (N=43)	ADVERTISING AND PUBLICITY NCOs (N=29)
TRAINING CLUSTER (N=23)	TECHNICAL SCHOOL INSTRUCTORS (N=14)
TEST CLUSTER (N=86)	
	PRODUCTION MANAGEMENT PERSONNEL (N=79)
	CLASSIFICATION INTERVIEWERS (N=18)

TABLE 3

COMPARISON OF PERCENT OF TIME SPENT PERFORMING DUTIES BETWEEN SURVEYS OF AFS 982X0 DENTAL LABORATORY PERSONNEL (1-24 MONTHS ACTIVE FEDERAL MILITARY SERVICE)

DUTIES	APR 74 SURVEY	JUN 78 SURVEY
ORGANIZING AND PLANNING	1	1
DIRECTING AND IMPLEMENTING	2	2
INSPECTING AND EVALUATING	*	1
TRAINING	1	1
PERFORMING ADMINISTRATIVE AND SUPPLY TASKS	3	1
PERFORMING GENERAL LABORATORY TASKS	48	53
FABRICATING AND REPAIRING COMPLETE DENTURES	13	10
FABRICATING AND REPAIRING REMOVABLE PARTIAL DENTURES	16	10
FABRICATING CROWNS INLAYS AND FIXED PARTIAL DENTURES	9	11
FABRICATING PORCELAIN PRODUCTS	1	7
FABRICATING AND REPAIRING ORTHODONTIC APPLIANCES	5	3
FABRICATING SPECIAL PROSTHESES	1	*

* INDICATES LESS THAN 1 PERCENT

263

TABLE 4

COMPARISON OF PERCENT OF MEMBERS PERFORMING TASKS BETWEEN SURVEYS OF AFS 982X0 DENTAL LABORATORY PERSONNEL

(1-24 MONTHS ACTIVE FEDERAL MILITARY SERVICE)

TASKS	APR 74 SURVEY	TASK DIFFICULTY	JUN 78 SURVEY	TASK DIFFICULTY
FABRICATE CRANIAL IMPLANTS	5	7.38	2	7.33
SOLDER METAL FRAMEWORKS USING OXYGEN-GAS TORCHES	12	6.88	7	5.80
DRAFT BUDGET OR FINANCIAL REQUIREMENTS	17	6.82	12	6.86
DEVELOP TESTS	9	6.72	5	6.85
ATTACH WIRES TO MODELS FOR ORTHODONTIC APPLIANCES	32	6.72	38	5.39
PLAN LAYOUT OF FACILITIES	13	6.58	9	6.85
FABRICATE PECTRIS-EXCAVATRIM IMPLANTS	5	6.51	7	7.46
TRIM DIES FOR WAXING	33	6.50	24	6.49

TABLE 5

TASK DIFFICULTY INDEX OF DIFFICULT TASKS BETWEEN SURVEYS OF RECRUITER PERSONNEL

TASK	TASK INDEX MAR 73 SURVEY	 MAY 78 SURVEY
SET UP PRODUCTION PLANS TO MEET FUTURE REQUIREMENTS	5.24	7.73
SUPERVISE AFSC 99120 PERSONNEL	4.96	8.04
CLARIFY, VALIDATE, AND OVERCOME PROSPECTS OBJECTIONS TO AIR FORCE ENLISTMENT	4.95	7.29
DESIGN OR IMPROVE WORK METHODS AND PROCEDURES	4.80	7.51
PLAN OR CONDUCT TRAINING CONFERENCES OR MEETINGS	4.67	6.61
CONDUCT FOLLOW-ON TRAINING OF NEWLY ASSIGNED PERSONNEL	5.46	7.36
ACCOMPLISH WAIVERS, REQUEST FOR APPROVAL OF ENLISTMENT FORMS	4.47	6.31
DETERMINE PRIMARY INTEREST OF PROSPECTS	4.47	6.08
IDENTIFY AND CONDUCT SPECIAL ASSISTANCE TRAINING FOR RECRUITERS OR RECRUITER STUDENTS	4.45	7.11

TABLE 6

COMPARISON OF PERCENT OF MEMBERS PERFORMING TASKS BETWEEN SURVEYS OF 5-SKILL LEVEL INVENTORY MANAGEMENT PERSONNEL

TASKS	DEC 68 SURVEY	OCT 72 SURVEY	JUL 78 SURVEY
MAINTAIN SUSPENSE FILES	42	18	45
COUNT PROPERTY	26	20	18
PREPARE ISSUE DOCUMENTS	26	19	19
COMPARE PHYSICAL COUNTS OF PROPERTY WITH STOCK RECORD BALANCES	18	16	16
PLACE LOCATION SYMBOLS ON STORAGE FACILITIES	11	4	5
PREPARE TURN-IN DOCUMENTS	10	11	18
ESTABLISH BENCH STOCKS	9	6	11
OPERATE REMOTE KEYBOARD UNITS	15	25	39

The Collection and Prediction of Training Emphasis Ratings for Curriculum Development

by

Hendrick W. Ruck
Nancy A. Thompson
David C. Thomson

Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper are those of the authors and are not necessarily those of the United States Air Force.

One of the most difficult questions that arises in occupational curriculum design is, "What should the training content be?" This question, in the business of Air Force vocational training, could be further reduced to the fundamental questions of "Which occupational tasks should be included in the curriculum?" and "How do those tasks translate into specific skills and knowledge?" The purpose of this paper is to address the first of the fundamental questions; i.e., the selection of tasks for training.

The Air Force Human Resources Laboratory has been conducting extensive research in the training requirements area. The initial concepts and theory guiding the research were first proposed by Christal (1970), who suggested that boards of expert judges could be used to study information about tasks that are hypothesized to be related to the training decision. The experts could then evaluate those tasks in terms of the appropriateness for inclusion in curricula. He further suggested that the mathematical technique of policy capturing be applied to the judges' decisions so that the policy of the judges could be applied to additional tasks. This approach would reduce the necessity of expert judgment in task selection for each task and would assure more consistent decisions since the mathematical model of the experts' decisions could be used instead of additional judgments by the experts.

These initial suggestions have been studied in a stream of research on task training factors. Mial and Christal (1974) developed a number of task training factors and were able to predict judges' mean rank ordering of tasks for priority in training using a four-factor regression equation ($R=.88$; $P<.001$). Their research was conducted using the Medical Service specialty. Mead (1975) has presented additional evidence as to the utility of the policy-capturing approach. He performed a similar study to that of Mial and Christal using a different specialty (Law Enforcement) and met with similar success in predicting training priorities. Mead successfully used both mean rankings of training priority and mean ratings of training priority in his research. These studies suggested that a promising link between Instructional System

Development (ISD) theory, occupational survey data, and curriculum design could be further developed.

Stacy, Thompson, and Thomson (1977) presented a paper last year at the Military Testing Association Conference outlining preliminary results of task factor data collection, training emphasis prediction, and task-anchored scaling. Stacy found that the task training factors could be collected reliably using standard occupational survey techniques. He also reported success in using the policy-capturing approach for a number of specialties. The present paper will discuss the results of policy-capturing research on 13 Air Force specialties, the similarities and differences in policies for different specialties, and the implications of the research for Instructional System Development (ISD). A separate paper is being presented at this conference by Squadron Leader David C. Thomson (Thomson and Goody, 1978) documenting the results of the task anchored scaling research.

Method

Research conducted prior to Stacy, et al (1977) focused on two specialties. This study was designed to test the generalizability of earlier findings. Therefore, 13 of the 14 specialties studied by Stacy, et al, were selected for this study (Table 1). The specialties were selected so that occupational survey data and job inventories were current, initial skill courses were mandatory for entry into the specialties, and all four aptitude areas used in Air Force job placement (Mechanical, Electrical, General, and Administrative) were represented. As a result of the operational occupational surveys conducted by the AF Occupational Measurement Center, data were available on percent members performing each task, an index of percent time spent on each task, the learning difficulty of each task, and the average grade of members performing each task. Additional data that were collected for the study included: (a) field recommended training emphasis for each task, (b) present school emphasis for each task, (c) probable consequences of inadequate performance for each task, and (d) delay tolerance for each task. The learning difficulty task factor used in this study was collected using a nine-point relative scale. However, the other two factors (consequences of inadequate performance, task delay tolerance) were collected using nine-point scales (Stacy, Thompson & Thomson; 1977) that were verbally anchored and did not require relative task comparisons. These factors have been described previously (Stacy, et al, 1977); however, the training emphasis scale will be described again in this paper because of its importance.

The field recommended training emphasis scale was developed as the criterion. It was expected to yield equivalent information to the mean rank orderings of training priority as used by Mial and Christal (1974), since Mead (1975) demonstrated the equivalency between rankings and ratings for training priority. The field recommended training emphasis scale is a nine-point scale ranging from "Extremely Little" to "Extremely Heavy." Senior NCOs serving in operational units in each specialty are

Table 1
AFSC Aptitude Areas and Raters

AFSC	Title	Aptitude Area	Number of Respondents/Raters				
			Members Total	Training Emphasis	Consequences	Delay	Difficulty
293X3	Radio Operator	A	1468	224	45	50	78
304X0	Radio Relay Equipment	E	1573	215	35	50	89
304X4	Ground Radio Communication Equipment	E	2351	335	60	58	122
328X3	Electronic Warfare	E	1223	306	46	47	43
472X2	General Purpose Vehicle Mechanic	M	3338	291	33	34	127
552X5	Plumbing Specialist	M	964	143	82	62	116
651X0	Procurement Specialist	A	979	320	61	63	101
672X1	General Accounting	A	596	85	55	55	86
672X2	Disbursement Accounting Specialist	A	1352	149	65	65	86
902X0	Medical Services	G	2198	380	93	95	58
906X0	Medical Administration	G	2356	300	105	104	78
911X0	Physiological Training	G	408	79	30	30	86
981X0	Dental Specialist	G	1856	89	65	47	45
Total			20662	4220	1096	1098	1451

asked to (a) check each task for which formal training (school or on-the-job training (OJT)) is recommended for first-term airmen, and (b) rate each of the tasks that were checked using the nine-point scale. The training emphasis scale is normally treated in data reduction and analysis as a 10-point scale since the absence of a check mark is treated as zero. This differs from other ISD task factors used in the Air Force occupational survey program since every task is normally considered to possess some amount of each ISD factor. That is, for example, no task would be expected to have zero learning difficulty. Similarly, no task would have zero consequences of inadequate performance or delay tolerance. Tasks could, however, have zero field recommended training emphasis.

The field recommended training emphasis scale has been intensively researched. It has been collected in the research mode for 19 specialties and in the operational mode for an additional 21 specialties. Table 2 lists the AF specialties and associated interrater agreement data for the field recommended training emphasis data collected to date. The median interrater agreement coefficient is .95. Analyses of rater agreement data suggest that a minimum of 40 raters should be used to provide reliable results for the recommended training emphasis scale.

The validation of field recommended training emphasis was performed using policy capturing (Christal, 1968). Policy capturing requires that a multiple regression model be developed in an attempt to "capture" the policy of the judges in their ratings or rankings. Basically, it is the development of explanatory and predictive regression models. The policy model that was developed to predict field recommended training emphasis included three task factors and three related job factors, together with squares of the factors. The task factors in the model were learning difficulty, probable consequences of inadequate performance, and task delay tolerance. The job-related factors were percent members performing in the first assignment, an index of percent time spent by members in their first assignment, and the average weighted grade of members performing. Since each factor was squared to address expected curvilinear relationships, a twelve variable regression equation was generated for each Air Force specialty.

The ISD literature has often been interpreted as suggesting that there is one correct way to combine task and job factor information in order to derive training requirements. This hypothesis was tested by analyzing the regression equations (Ward, 1963; Gott, 1978) for each specialty to determine whether, in fact, different policies as expressed in the policy equation exist across specialties, or, as one might expect, there is one universal equation (or combination rule). The analyses required to test the hypothesis can be conducted using a hierarchical grouping algorithm which tests similar regression equations for homogeneity of weights.

Table 2
Descriptive Statistics of Training Emphasis Ratings

AFSC	Title	Avg Mean	SD	R_{kk}*	No. of Raters
111X0	Defensive Aerial Gunner	3.49	1.94	.94	43
293X3	Ground Radio Operator	2.26	1.44	.92	189
303XX	Aircraft Control and Warning Radar	3.08	1.76	.96	52
304X0	Radio Relay Equipment	2.38	1.50	.94	199
304X4	Ground Radio Communication Equipment	1.82	1.15	.90	315
307X0	Telecommunication Systems Control	2.87	2.11	.96	75
321XX	Defensive Fire Control Systems	2.48	2.09	.97	50
328X3	Electronic Warfare Systems	1.02	1.25	.94	248
341XX	Training Devices	1.38	1.15	.89	46
361X0	Outside Wire and Antenna Maintenance	3.53	1.67	.94	40
423X1	Aircraft Environmental Systems	2.63	1.34	.91	137
423X4	Aircraft Pneudraulic Systems	2.89	1.55	.93	282
427X2	Nondestruction Inspection	3.89	2.29	.98	178
427X5	Airframe Repair	3.48	2.16	.97	267
443X0	Missile Maintenance (LGM 25-Titan)	3.91	2.30	.97	53
462X0	Aircraft Armament Systems	2.72	1.81	.96	186
463X0	Nuclear Weapons	2.05	2.29	.90	40
472X2	General Purpose Vehicle Maintenance	2.39	2.30	.98	243
472XX	Vehicle Maintenance	3.42	1.75	.95	49
542X2	Electronic Power Production	3.85	1.54	.93	40
552X5	Plumbing	3.32	1.62	.95	125
555X0	Programs and Work Control	2.61	1.48	.92	54
571X0	Fire Protection	3.55	1.87	.95	51
601X4	Packaging	1.42	1.72	.97	17
602XX	Passenger and Freight	2.75	1.58	.92	26
622X1	Diet Therapy	3.68	1.82	.95	47
631X0	Fuel	3.22	1.92	.95	277
645X0	Inventory Management	1.77	1.55	.92	12
645X1	Materiel	1.27	1.37	.92	15

Table 2 (Continued)
Descriptive Statistics of Training Emphasis Ratings

AFSC	Title	Avg Mean	SD	R_{kk}*	No. of Raters
645X2	Supply	1.31	2.25	.99	5
651X0	Procurement	2.85	1.63	.94	295
672X1	General Accounting	1.29	1.41	.95	73
672X2	Disbursement Accounting	1.21	1.37	.95	131
902X0	Medical Service	3.44	1.72	.93	302
904X0	Medical Laboratory	2.91	1.83	.95	46
906X0	Medical Administrative	1.71	1.19	.91	270
907X0	Environmental Health	3.25	1.82	.95	64
911X0	Aerospace Physiology	2.59	2.00	.96	68
981X0	Dental	2.33	2.06	.97	85
982X0	Dental Laboratory	2.84	2.02	.97	23

*Rater agreement indices for a sample of 40 raters as estimated by the Spearman Brown formula.

Analysis of Policy Equations

The results of the grouping analysis of the 13 policy equations are highlighted in Table 3. Notice that, if the regression equation derived for each of the 13 specialties was used to predict for that specialty, the overall predictive efficiency would be quite high (.86). On the other hand, predictive efficiency using a single averaged equation for each of the specialties would result in unacceptably low predictive efficiency of .56. As a result of the analysis, a compromise solution appears to be one which uses one equation (Policy A) for eight specialties, and a second equation (Policy B) for the remaining five specialties. The equation used in Policy A yields an R-squared of .72; however, the Policy B equation has an R-squared of .64. This suggests that the specialties in Policy B are not as predictable (using the ISD factors) as those in Policy A.

Table 3
Grouping of Training Priority Policy Equations

	Number of Equations	Overall Predictive Efficiency (R^2)
Maximum	13	.86
Optimal	2	.73
Minimum	1	.56

Additional analyses of the differences between the two policies were performed in an attempt to isolate characteristics of specialties in each policy. The specialties in Policy B differed from those in Policy A in that Policy B specialties were measured with job inventories that had significantly more tasks than in A ($\overline{X}_A = 425$, $\overline{X}_B = 951$; $t = 3.81$, $df = 11$, $p<.01$) and Policy B specialties included significantly more job types than Policy A ($\overline{X}_A = 27.1$, $\overline{X}_B = 47.6$, $t = -2.69$, $df = 11$, $p<.05$). It is important to note here that no significant relationship was found between number of tasks in a job inventory and number of job types identified in an occupational analysis ($r = .29$, ns).

Analysis of Interrater Agreement

Although the complex relationships among recommended training emphasis and the ISD factors will not be more fully developed in this presentation, one may conclude that there is no single method of combining ISD factor data to arrive at training emphasis for all specialties. This conclusion applies if Air Force specialties are considered the unit of analysis; however, the conclusion has not been tested, and may not hold up if the unit of analysis is changed to job groups within specialties rather than specialties. Nevertheless, the finding is significant, since most technical training in the Air Force is developed for specialties and not for job groups.

After determining that there were at least two different policies that could be used to predict training emphasis, and that the policies differed in predictive efficiency and type of specialty, further analyses of interrater agreement were conducted. Although no difference in interrater agreement was found among the specialties in each policy group, interrater agreement was found to be moderately correlated with predictive efficiency. That is, the correlation between R-squared for each specialty and interrater agreement (R_{11}) on training emphasis is significant ($r = .61$, $p<.05$).

Table 2 displays the interrater agreement values adjusted for 40 raters. Using a conservative cutoff of .91 for acceptable interrater agreement, one can see that five (or 12.5 percent) of the specialties do not meet the cutoff. This analysis leads to the conclusion that the recommended training emphasis data can be reliably collected in at least 80 percent of the Air Force specialties.

Complex Specialties

The training emphasis research has resulted in a criterion that may be collected and used in decision making in a large number of specialties. Two types of problem specialties have been identified in the research. First, there are specialties with low interrater agreement. Second, there are specialties for which predictability of recommended training emphasis is not as high as is necessary for practical prediction. These complex specialties are being investigated further. The new research will attempt to (a) determine whether additional factors may be useful for predicting recommended training emphasis in complex specialties through the development of additional task factor scales, (b) examine the complex specialties for common characteristics, (c) determine optimal data displays for complex specialties for training decision makers, (d) determine which specialty characteristics are associated with low interrater agreement and poor predictability.

Discussion

The task training factor research stream has produced significant results. First, ISD task and job factors have been identified and scales developed to measure them. Second, the field recommended training emphasis scale has been developed as a criterion. The training emphasis scale has been shown to be reliable, through interrater agreement analyses in 40 specialties, and valid, through policy capturing and policy grouping in 13 specialties. Third, no single way of combining ISD factors for training decisions was found to be appropriate for all specialties.

The initial objective of the research was to discover combination rules that may be applied to ISD factor data for selecting tasks for training. This has been done. However, the rules differ for different groups of specialties. The criterion used in the research has been found to be both reliable and valid. Furthermore, only a moderate number of raters (about 40) is required to provide stable data. These findings have led to the unexpected conclusion that the criterion should be collected and not predicted. It is important to note that recommended training emphasis ratings, although useful for most Air Force specialties, will not always be immediately usable. In particular, complex specialties appear to require additional study to enhance understanding of low predictive efficiency and poor interrater agreement.

As a result of this research, it has been recommended that supervisory ratings of formal training emphasis be collected routinely in the Air Force Occupational Survey Program. Further, it has been recommended that routine collection of the task factors Consequences of Inadequate Performance and Task Delay Tolerance be discontinued since recommended training emphasis ratings include consideration of those factors. In cases where more than one Air Force specialty would be included in a single job inventory, it is recommended that separate ratings be collected for each specialty and that those ratings be analyzed and presented for each of the specialties. Finally it has been recommended that the training emphasis data be presented using new modularized CODAP (Thew & Weissmueller, 1978) programs that allow a merging of training documentation, such as Specialty Training Standards (STS) or Plans of Instruction (POI) with job inventory tasks. This merging of job inventory tasks and training documents provides a simple and reliable method of displaying occupational survey data within the context that training personnel are most familiar. The Appendix displays an example of the output.

The research leading to the conclusions and recommendations has been difficult and complex. However, the available technology for using occupational survey data for training decisions has been considerably expanded and implementation of the results would provide a much stronger basis for making training decisions than is currently available.

References

Christal, R.E. Implications of Air Force occupational research for curriculum design. In B. B. Smith & J. Moss, Jr. (Eds.), *Report of a seminar: Process and Techniques of vocational curriculum development*. Minnesota Research Coordination Unit for Vocational Education, University of Minnesota, Minneapolis, MN, April 1970, 27-61.

Christal, R.E. JAN: A technique for analyzing group judgment. *Journal of Experimental Education*, 1968, 36(4), 24-27.

Gott, C.D. *HIER-GRP: a computer program for the hierarchical grouping of regression equations*. AFHRL-TR-78-14. Brooks Air Force Base, TX: Computational Sciences Division, Air Force Human Resources Laboratory, June 1978.

Mead, D.F. *Determining training priorities for job tasks*. Paper presented to 17th Annual Conference of the Military Testing Association, U.S. Army, Indianapolis, IN, 16-19 September 1975.

Mial, R.P. & Christal, R.E. The determination of training priority for vocational tasks. *Proceedings, Psychology in the Air Force Symposium*. USAF Academy, April 1974, 29-33.

Stacy, W.J., Thompson, N.A. & Thomson, D.C. *Occupational task factors for instructional systems development*. Paper presented to 19th Annual Conference of the Military Testing Association, USAF, San Antonio, TX, 17-21 October 1977.

Thew, M.C. & Weissmuller, J.J. *CODAP: a modular approach to occupational analysis*. Paper presented to 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 October - 3 November 1978.

Thompson, D.C. & Goody, K. *Benchmark scales for collecting task training factor data*. Paper presented to 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 October-3 November 1978.

Ward, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963, 58, 236-244.

APPENDIX

1. This appendix contains sample computer output that is being recommended for use by training managers and curriculum developers. The printout is in two parts; the first part (pp 12-14) is an executive summary, and the second part (pp 15-16) provides detailed Occupational Survey (OS) data that have been matched with Specialty Training Standard (STS) items.

2. Four columns are on each of the subsequent printouts. Column 1 displays the number of OS tasks that have been matched to each of the STS items. The data displayed in column 2 are the recommended training emphasis ratings collected from AFS 293X3 (Ground Radio Operator) field supervisors. Column 3 includes the percent of job incumbents with 2-24 months military service in the 293X3 career ladder who perform the tasks. Finally, Column 4 shows the learning difficulty for each task.

3. The executive summary has been designed to aggregate OS task data to STS item level. Pages 12-14 display STS items in original STS sequence. Mean values for each of the three OS task factors for the tasks that were matched to the STS item are displayed in columns 2-4. Note that in the case where no tasks have been matched with an STS item, the values in the adjacent columns are zero. Also, it is possible to show the same STS items and OS data with the STS items arranged in descending order on field supervisors' recommended training emphasis. This display gives a powerful overview of the data and is expected to be quite useful to training managers.

4. The remaining two pages (pages 15-16) provide samples of detailed OS data in the STS framework. STS items are printed between dashed lines and OS tasks (and associated data) are shown immediately below the items. Tasks are listed in order of training priority for introductory airmen within each STS category. The matching between STS and OS tasks was performed and reviewed by course personnel. Note that tasks may be mapped into as many STS items as required and that both STS items and OS tasks may have no counterparts.

STS - OSR EXECUTIVE SUMMARY FOR AFS 293X3 STS05A

293X3 - RADIO OPERATOR (STUDY 6347)

STS NUM	STS ITEM TITLE	NUM TSK (-)	TNG EMP (F)	MEM -24 (M)	LRN DIF (F)
1	CAREER PROGRESSION	0	.00	.0	.00
1A	AIRMAN COMMUNICATIONS OPERATIONS CAREER FIELD	1	1.34	5.5	4.04
1B	PROGRESSION IN CAREER LADDER 293X3	1	1.34	5.5	4.04
1C	CAREER MOTIVATION	1	1.34	5.5	4.04
2	SECURITY	0	.00	.0	.00
2A	COMMUNICATIONS SECURITY (COMSEC)	7	3.80	29.1	3.90
2B	OPERATIONS SECURITY (OPSEC)	1	1.75	6.2	4.35
3	SUPERVISION AND TRAINING	0	.00	.0	.00
3A	SUPERVISION	6	1.18	10.6	4.16
3A1	EVALUATE PERFORMANCE OF PERSONNEL AND COMPLETE APPROPRIATE RATING FORMS	3	1.09	4.3	4.56
3A2	ORIENT NEWLY ASSIGNED PERSONNEL TO THE ORGANIZATION AND MISSION OF THE UNIT.	4	1.52	9.1	3.94
3A3	INITIATE CORRESPONDENCE & SOPS CONCERNING RADIO ASSIGNMENTS	11	1.85	12.5	4.40
3A4	ESTABLISH PRIORITIES AND SCHEDULE WORK ASSIGNMENTS	14	1.55	10.0	4.26
3A5	SUPERVISE RADIO OPERATING ACTIVITIES	37	1.50	8.4	4.22
3B	TRAINING	5	1.24	10.8	4.08
3B1	EVALUATE PERSONNEL TRAINING NEEDS	3	1.23	3.7	3.97
3B2	RECOMMEND PERSONNEL FOR TRAINING	6	1.12	9.0	4.06
3B3	PLAN, CONDUCT AND SUPERVISE OJT	13	1.43	9.0	4.34
3B4	PREPARE JOB PROFICIENCY GUIDES	3	1.71	5.1	4.66
3B5	MOTIVATE TRAINERS AND TRAINEES	2	.95	3.5	3.85
3B6	COUNSEL TRAINEES ON TRAINING PROGRAMS	1	3.36	14.3	4.72
3B7	MONITOR EFFECTIVENESS OF OJT	5	1.76	7.3	4.23
3B8	MAINTAIN TRAINING RECORDS	4	1.49	11.6	4.08
3B9	EVALUATE EFFECTIVENESS OF TRAINING PROGRAMS	3	.84	5.0	4.13
4	COMMUNICATIONS AGENCIES, SYSTEMS/FACILITIES, PUBLICATIONS & RECORDS	0	.00	.0	.00
4A	COMMUNICATIONS AGENCIES	0	.00	.0	.00
4B	COMMUNICATIONS SYSTEMS/FACILITIES	12	2.83	14.0	3.75
4C	COMMUNICATIONS PUBLICATIONS	7	2.23	13.4	4.11
4D	COMMUNICATIONS RECORDS	24	2.82	25.0	3.24
5	OPERATIONAL INSPECTIONS	0	.00	.0	.00
5A	PERFORM INSPECTIONS OF ON-LINE AND OFF-LINE RADIO EQUIPMENT	2	3.74	20.5	3.97
5B	DETECT AND REPORT EQUIPMENT PROBLEMS TO THE APPROPRIATE MAINTENANCE SUPPORT ACTIVITY	2	2.47	15.9	3.36
6	EQUIPMENT TUNING AND OPERATION	0	.00	.0	.00
6A	APPLY SAFETY PRECAUTIONS PERTAINING TO ELECTRICAL EQUIPMENT	4	2.87	10.8	3.97
6B	TUNE AND ADJUST RECEIVER/TRANSCEIVER CONTROLS	17	4.56	32.3	3.62
6B1	TO CALIBRATE, USING REMOTE SIGNALS OF A KNOWN FREQUENCY	0	.00	.0	.00
6B2	ACCORDING TO DESIRED TYPE OF EMMISSION	0	.00	.0	.00
6B3	TO MAINTAIN A READABLE SIGNAL	1	3.95	21.6	3.55
6B4	TO COUNTER THE EFFECTS OF MEACONING, INTRUSION, JAMMING OR INTERFERENCE (MIJI)	4	3.66	16.9	4.01
6C	TUNE AND ADJUST TRANSMITTER/TRANSCEIVER CONTROLS	13	4.73	35.3	3.67
6C1	TO THE DESIRED EMISSION	0	.00	.0	.00

STS - OSR EXECUTIVE SUMMARY FOR AFS 293X3 STSOSR

STS NUM	STS ITEM TITLE	NUM TSK (-)	ING EMP (E)	MEM -24 (M)	LRN DIF (F)
6C2	FOR APPROPRIATE OUTPUT POWER	0	.00	.0	.00
6C3	FOR PROPER ANTENNA LOADING	1	3.91	12.5	3.44
6D	USE CONTROL CONSOLE TO	10	4.33	25.4	3.52
6D1	SELECT TRANSMITTER/RECEIVER FREQUENCY	0	.00	.0	.00
6D2	SELECT ANTENNA SYSTEMS	0	.00	.0	.00
6D3	TUNE REMOTELY LOCATED TRANSMITTERS/RECEIVERS	0	.00	.0	.00
6D4	MONITOR OPERATING FREQUENCIES	0	.00	.0	.00
6D5	SELECT BACK-UP TRANSMITTERS/RECEIVERS	0	.00	.0	.00
6D6	ORIENT AND SELECT ANTENNA AZIMUTH	2	4.33	43.8	3.11
6D7	CONFIGURE EQUIPMENT FOR BACK TO BACK RELAY	1	4.13	22.3	3.74
6D8	PROVIDE PATCHING SERVICE USING SEMI-AUTOMATIC OR MANUAL TECHNIQUES (VOICE, RTTY, DIGITAL)	3	4.05	23.1	3.68
6D9	INITIATE AND REPLY TO CALLS ON ASSOCIATED TELEPHONES AND "HOTLINES"	0	.00	.0	.00
7	WAVE CREATION/PROPAGATION AND COLLECTION	0	.00	.0	.00
7A	WAVE CREATION/PROPAGATION	4	3.13	14.5	3.92
7B	COLLECTION	0	.00	.0	.00
7B1	TYPES AND FUNCTIONS OF HF GROUND STATION ANTENNA SYSTEMS	0	.00	.0	.00
7B2	ERECT OR INSTALL MOBILE TYPE ANTENNAS UNDER FIELD CONDITIONS	2	3.24	14.8	4.32
8	TRANSMITTING AND RECEIVING SKILLS (VOICE)	0	.00	.0	.00
8A	MAKE VOICE TRANSMISSIONS	0	.00	.0	.00
8B	RECEIVE AND TRANSCRIBE VOICE TRANSMISSIONS WITH A TYPEWRITER	1	5.12	44.7	3.62
9	OPERATING PROCEDURES (VOICE)	0	.00	.0	.00
9A	MAINTAIN WATCH ON ASSIGNED FREQUENCIES & USE THE PRESCRIBED	4	4.38	47.4	3.41
9A1	PHONETIC ALPHABET	1	5.97	73.6	3.62
9A2	PROWORDS/PROSIGNS	1	5.97	73.6	3.62
9A3	OPERATING SIGNALS	0	.00	.0	.00
9A4	HANDPRINTING TECHNIQUES	0	.00	.0	.00
9A5	UNIVERSAL TIME AND TIME ZONE CONVERSION	0	.00	.0	.00
9A6	MICROPHONE TECHNIQUES	0	.00	.0	.00
9A7	TESTING SIGNALS AND PROCEDURES	1	5.10	48.7	3.19
9A8	BROADCAST PROCEDURES	0	.00	.0	.00
9A9	MESSAGE FORMATS	0	.00	.0	.00
9A10	LOGKEEPING PROCEDURES	3	4.16	46.3	3.14
9A11	CALLING AND ANSWERING PROCEDURES	4	5.21	53.0	3.59
9A12	MESSAGE HANDLING PROCEDURES	7	4.46	35.3	3.66
9A13	PHONE PATCH PROCEDURES	2	5.70	58.4	3.56
9A14	RELAY AND ROUTING PROCEDURES	4	3.98	32.4	3.53
9A15	TRAFFIC SERVICING PROCEDURES	0	.00	.0	.00
9A16	DIRECTION FINDING (DF) PROCEDURES	3	3.36	25.9	3.78
9A17	RECORDING DEVICES AND PROCEDURES	2	3.68	30.4	2.58
9A18	UNUSUAL INCIDENT/INTERFERENCE OR CIRCUIT CONDITION REPORTING PROCEDURES	5	3.04	19.9	4.17
9A19	NET CONTROL PROCEDURES	3	4.10	49.6	3.05
9B	USE APPROVED GROUND/AIR RADIOTELEPHONE	5	4.67	39.9	3.90
9B1	ICAO PROCEDURES	4	5.01	29.5	3.65
9B2	GUARD STATION PROCEDURES	0	.00	.0	.00

```
STS - DSR EXECUTIVE SUMMARY FOR AFS 293X3                                  STSDSR

                                                                 NUM        TNG         MEM        LRN
STS   STS ITEM TITLE                                             TSK        EMP         -29        DIF
NUM                                                              (-)        (F)         (M)        (F)

9B3   AIRCRAFT EMERGENCY PROCEDURES                                3       4.46        37.2       4.05
9B4   TRANSMIT/RECEIVE WEATHER INFORMATION USING STANDARD WEATHER  2       4.16        23.3       3.45
       SYMBOLS AND ABBREVIATIONS
       TASKS NOT REFERENCED                                      159       1.56         2.1       3.15
```

```
STS - OSR EXECUTIVE SUMMARY FOR AFS 293X1                           5150 H
                                                                    TNG        MEM       LRN
                                                                    EMP        -24       DIF
D TSK          TITLES                                               (E)        (M)       (F)
---------------------------------------------------------------------------------------------
***************************************************************
343   INITIATE CORRESPONDENCE & SOPS CONCERNING RADIO ASSIGNMENTS
***************************************************************

E 128 TYPE RECORDS, REPORTS, OR FORMS                               3.00       39.1      3.95
E 127 TYPE CORRESPONDENCE                                           2.88       20.9      3.94
A  22 PLAN RADIO OPERATIONAL SUPPORT FOR EXERCISES OR               2.05        9.2      4.96
        SPECIAL MISSIONS
A   7 DEVELOP OPERATORS' CHECKLISTS                                 2.02       10.3      4.52
B  27 DIRECT IMPLEMENTATION OF EMERGENCY PROCEDURES TO              2.00       12.1      4.76
        SUPPORT DISASTER OR CONTINGENCY PLANS
C  99 INTERPRET POLICIES OR DIRECTIVES FOR SUBORDINATES             1.65        7.7      4.22
C  65 EVALUATE COMPLIANCE WITH WORK STANDARDS OR                    1.56        7.3      4.96
        OPERATING PROCEDURES
A  11 DEVELOP RADIO OPERATIONS COMMUNICATIONS OPERATING             1.50       12.5      4.93
        INSTRUCTIONS (COI)
A  21 PLAN OR SCHEDULE WORK ASSIGNMENTS OR SHIFT                    1.44        9.5      3.62
        SCHEDULES
A  19 ESTABLISH WORK CONTROLS OR PERFORMANCE STANDARDS              1.39        7.7      4.63
B  33 DRAFT, EDIT, OR REVIEW CORRESPONDENCE                          .92        6.6      4.44

***************************************************************
344   ESTABLISH PRIORITIES AND SCHEDULE WORK ASSIGNMENTS
***************************************************************

B  24 DIRECT OPERATIONS OF GROUND RADIO STATIONS AND                2.48       12.8      4.55
        ASSOCIATED EQUIPMENT
A  22 PLAN RADIO OPERATIONAL SUPPORT FOR EXERCISES OR               2.05        9.2      4.96
        SPECIAL MISSIONS
A   7 DEVELOP OPERATORS' CHECKLISTS                                 2.02       10.3      4.52
A   9 DEVELOP OR IMPROVE WORK METHODS OR PROCEDURES                 2.00       20.1      4.09
B  27 DIRECT IMPLEMENTATION OF EMERGENCY PROCEDURES TO              2.00       12.1      4.76
        SUPPORT DISASTER OR CONTINGENCY PLANS
A  15 ESTABLISH WORK PRIORITIES                                     1.89       13.2      4.02
C  65 EVALUATE COMPLIANCE WITH WORK STANDARDS OR                    1.56        7.3      4.96
        OPERATING PROCEDURES
A  21 PLAN OR SCHEDULE WORK ASSIGNMENTS OR SHIFT                    1.44        9.5      3.62
        SCHEDULES
A  19 ESTABLISH WORK CONTROLS OR PERFORMANCE STANDARDS              1.39        7.7      4.63
B  36 ESTABLISH PROCEDURES FOR CARE OR UTILIZATION OF WORK          1.23        6.6      4.02
        SPACE, EQUIPMENT, OR SUPPLIES
B  31 DISPATCH MOBILE RADIO UNITS                                   1.19        7.0      2.95
A   6 DETERMINE REQUIREMENTS FOR EQUIPMENT OR SUPPLIES              1.09       12.8      4.09
C  68 EVALUATE PROCEDURES FOR STORAGE, INVENTORY, OR                 .83        4.4      3.99
        INSPECTION OF PROPERTY ITEMS
A   5 DETERMINE PERSONNEL REQUIREMENTS                               .60        7.7      4.35
```

STS - OSR EXECUTIVE SUMMARY FOR AFS 293X0 STSOSR

D TSK	TITLES	TNG EMP (F)	MEM -24 (M)	TSK DIF (F)
G 201	PROCESS REQUESTS FOR ASSISTANCE, INFORMATION, OR INSTRUCTIONS FROM AIRCRAFT IN FLIGHT	5.63	39.6	3.96
G 100	IDENTIFY INCOMING CALLS USING CALL SIGN LIST	4.90	49.6	3.36
G 107	MAINTAIN CURRENT CALL SIGN LISTS	4.32	49.1	3.40

9A12 MESSAGE HANDLING PROCEDURES

D TSK	TITLES	TNG EMP (F)	MEM -24 (M)	TSK DIF (F)
G 201	PROCESS REQUESTS FOR ASSISTANCE, INFORMATION, OR INSTRUCTIONS FROM AIRCRAFT IN FLIGHT	5.63	39.6	3.96
G 173	COORDINATE AIR-TO-GROUND TRAFFIC	5.23	43.6	4.04
G 203	RELAY COMMUNICATIONS TRAFFIC BETWEEN FIXED STATIONS AND AIRCRAFT	5.16	47.6	3.56
F 192	COORDINATE TRAFFIC WITH OTHER AGENCIES OR UNITS, SUCH AS AIR TRAFFIC CONTROL, OR AIRBORNE COMMAND POSTS	4.77	39.2	3.63
G 207	ROUTE OR REROUTE AIRCRAFT MOVEMENT MESSAGES	4.56	21.2	3.57
G 184	LIST TRAFFIC WITH NET CONTROL STATIONS	3.71	40.3	2.90
A 17	PLAN OR ESTABLISH PROCEDURES FOR ALTERNATE ROUTING OF TRAFFIC	2.16	15.6	3.78

9A13 PHONE PATCH PROCEDURES

D TSK	TITLES	TNG EMP (F)	MEM -24 (M)	TSK DIF (F)
G 186	MAKE PHONE PATCHES	5.78	77.3	3.16
G 201	PROCESS REQUESTS FOR ASSISTANCE, INFORMATION, OR INSTRUCTIONS FROM AIRCRAFT IN FLIGHT	5.63	39.6	3.96

9A14 RELAY AND ROUTING PROCEDURES

D TSK	TITLES	TNG EMP (F)	MEM -24 (M)	TSK DIF (F)
G 173	COORDINATE AIR-TO-GROUND TRAFFIC	5.23	43.6	4.04
G 203	RELAY COMMUNICATIONS TRAFFIC BETWEEN FIXED STATIONS AND AIRCRAFT	5.16	47.6	3.56
F 192	COORDINATE TRAFFIC WITH OTHER AGENCIES OR UNITS, SUCH AS AIR TRAFFIC CONTROL, OR AIRBORNE COMMAND POSTS	4.77	39.2	3.63
G 207	ROUTE OR REROUTE AIRCRAFT MOVEMENT MESSAGES	4.56	21.2	3.57
G 205	RELAY COMMUNICATIONS TRAFFIC BETWEEN MOBILE STATIONS AND AIRCRAFT	4.30	9.2	3.12
G 204	RELAY COMMUNICATIONS TRAFFIC BETWEEN FIXED STATIONS AND MOBILE STATIONS	4.22	30.6	3.26
G 188	MAKE TIME CHECKS	4.16	60.1	2.33
A 17	PLAN OR ESTABLISH PROCEDURES FOR ALTERNATE ROUTING OF TRAFFIC	2.16	15.6	3.78
A 9	COORDINATE WORK ACTIVITIES WITH OTHER UNITS OR SECTIONS	1.23	24.5	4.28

DATA BASE TO DETERMINATION OF TRAINING CONTENT:

A MANAGEABLE SOLUTION

Douglass Davis

The Chief of Naval Education and Training (CNET), Naval Air Station, Pensacola, Florida, is the Chief of Naval Operations (CNO) designee as the Navy's principal training agent. The CNET participates, inter alia, in the development and implementation of the most effective teaching and training systems and devices for optimal education and training. This paper will describe in some detail a CNET initiative which is bringing into a manageable focus the historical problem of determining the content of training programs within the agonizing limitations of existing, and even diminishing, resources.

Although a framework does exist for determining the training requirements of naval personnel, there is inconsistency among assigned roles of the Navy's three "Training Warfare Desks" (within the CNO). This situation undoubtedly springs from the fact that within the United States Navy there are three distinct communities: air, surface, and submarine. The distinctions are so prevalant that personnel within the employ of the Navy Department often speak of three "separate Navies." To illustrate the reality of this situation, one need only refer to the CNO instruction which delineates the functions of the three individual Training warfare Desks: OP-29, OP-39, and OP-59, for submarine, surface, and aviation manpower and training requirements, respectively.

A function of the Submarine Manpower and Training Requirements Division is the identification and establishment of training concepts and requirements; the corresponding function of the Surface Warfare Manpower and Training Requirements Division is the identification of requirements and the establishment of priorities for assigned training programs. The Aviation Manpower and Training Division, however, is tasked with developing requirements for aviation training courses of instruction conducted by the CNET and with exercising curriculum control and ensuring a continuum of training by coordinating the integration and standardization of flight, aviation ground and aviation technical training conducted by the Chief of Naval Education and Training

The clue to dealing with this disparity lies perhaps in the one common function among the three Training Warfare Desks: developing (or establishing) training requirements. The vehicle for systematically specifying requirements lies in the surface (OP-39) function of establishing priorities for assigned training programs. It is the implementation of the Instructional Systems Development (ISD) process which enables the CNET and the respective Manpower and Training Warfare Divisions to make possible the development (quantifiable statement) of requirements and the establishment of priorities within stated requirements.

An early product of the ISD process is a Job Task Inventory (JTI) or list of tasks which school or course graduates may reasonably be expected to perform in their fleet (or shore) assignments. It is the JTI which actually serves as a statement of training requirements and gives CNET and sponsors at the CNO echelon a data base from which to negotiate in the ultimate determination of training content. This process of negotiation of training requirements has been in progress since early March of this year (1978) following the critique of the Radioman (RM) "A" School proposed curriculum validation at the Service School Command, San Diego. This critique was attended by representatives from the RM Technical Advisor, the CNO rating advisor, and the Commanders in Chiefs, Atlantic and Pacific Fleets.

At this critique, primarily because of the attendees' inability to agree upon curriculum content, the concept of prioritization of JTI items (Job tasks) was introduced by CNET representatives. The plan, which has been recently carried out to completion, involved the forwarding of a CNET-developed JTI to CNO for subsequent distribution to the Rating Technical Advisor (COMNAVTELCOM), the Commander in Chief Atlantic Fleet (CINCLANTFLT), and the Commander in Chief Pacific Fleet (CINCPACFLT) and their type commanders (air, surface, and submarine). Each recipient of the JTI prioritized the list of tasks from most to least critical and forwarded the prioritized listing up the chain of command to CNO. In early October 1978, CNO forwarded the consolidated prioritized list of tasks to CNET as a formal statement of training requirements for RM "A" School (apprentice) trainees. The prioritization contains three sections: Priority A - Major Tasks identified as CRITICAL; Priority B - Tasks Identified as IMPORTANT; and Priority C - Tasks identified to be included if practicable, for example:

CATEGORY A: (CRITICAL)

Receive top secret material
Receive secret material
Process confidential material

CATEGORY B: (IMPORTANT)

Update crypto center files
Perform operator maintenance on TSEC/KW-26

CATEGORY C: (Include if practicable)

Inventory parts/tools/supplies

Upon receipt of the prioritized JTI, CNET has begun to study the requirements so stated in order to determine exactly how far down the list of tasks the Naval Education and Training Command can afford, within current assets, to successfully develop training programs to satisfy fleet and OPNAV expectations. For the first time, CNET is able to work from prioritized, approved lists of requirements. When resources have been exhausted, CNET can continue this cooperative endeavor with CNO to determine the placement of tasks which cannot be trained in the RM "A" School within the bounds of present numbers of student billets, school staff billets, equipments, and OPN/O&MN funding. Of course the CNO will have the option of reallocating resources to the RM "A" School or of supporting additional resource allocations in the outyears. Exercising this option may include the assignment of training tasks to On-the-job Training (OJT), to Self-Training Exportable Packages (STEPs), or to Rate Training Manuals and/or Career Correspondence Courses.

This venture, emanating from the data base created by application of the ISD process, enables all concerned to plan career training and to understand the rationale which determined the placement of training in a particular setting or at a particular point in the career of enlisted RMs. The dilemma of curriculum content will now begin to deminish as determination of curriculum content is removed from those who develop curricula and is placed in the hands of those who have actually been charged for many years with determining training requirements.

As one would well expect, Fleet recipients of course graduates find their jobs easier when course graduates have been trained to perform at identified, approved high levels of competency. Ideally, CNET would ensure that graduates meet or exceed the expectations of their supervisors. It is these expectations and the limitations of resources and student billets which have made for misunderstandings, illogically derived course content, and generally uncomfortable feelings among the Fleets who receive graduates and the command (CNET) that develops and administers training programs. Admittedly, there has been confusion concerning expected and actual performance of CNET course graduates. This situation has existed for several years primarily because of CNET's having been forced to remove "extraneous" material from courses and to train only to "need to know" in order to shorten courses whenever practicable to ease the impact of decreasing resources.

A follow-on to prioritization of training requirements and development of courses which reflect this prioritization is the development and refinement of a concept which, once implemented, will serve to preclude misunderstandings on the part of active users of course graduates. A Skills Profile (SP) will be developed for the purpose of "profiling" the job entry level for RM "A" (and ultimately all other) School graduates. The SP will enumerate the skills possessed by graduates and will be made available to all cognizant activities via the Catalogue of Navy Training Courses (CANTRAC) microfiche medium. Such a precise statement of capabilities to which a graduate has been trained will provide a definitive baseline against which job performance can be evaluated and from which a Fleet feedback system and a training readiness index can be implemented.

288

The cooperative CNO-CNET effort to provide a data base and the subsequent prioritization of training requirements which it supports will provide CNO sponsors data upon which to base decisions (the "who, what, when and where" of training), while making possible the realization of the actual CNET role: applying expertise in designing, developing, implementing, and evaluating (the "how" of training) courses which will, more than ever before, meet Fleet requirements.

REFERENCES

Davis, D. ISD Milestone: Radioman "A" Curriculum . . . Critique CAMPUS (The Navy Education and Training Monthly). Vol. VII No 5, May 1978, 4-5.

Department of the Navy. OPNAV Instruction 5430.48A. Washington, D.C. May 1977.

Department of the Navy. OPNAV Instruction 5450.194. Washington, D.C. Feb 1977.

NAVEDTRA 106A. Interservice Procedures for Instructional Systems Development. Aug 1975.

USING THE COMPUTER TO BUILD THE TASK INVENTORY

T. M. Ansbro

Career Development Group, Naval Education and Training
Program Development Center, Pensacola, Florida

At the "front end" of Instructional Systems Development (ISD) occupational data stockpiles, especially when data gathering is enthusiastically and thoroughly pursued. Some of the data gathering for Training Task Analysis in the Navy has so far produced thousands of tasks per rating, and there is no evidence to suggest a change in the trend. As the data-gathering techniques necessarily (and unavoidably) become more sophisticated and complex, the chances are increased that the data recorded will be sufficiently comprehensive to permit follow-on analysis to perform its design function in development of training curricula and materials, certainly throughout and hopefully far beyond any initial iteration of ISD. Except for technological change or significant adjustments in manpower management, there should be little need for more than a periodic augmentation to a data storage that has been assembled with a broad compass of retrieval strategies in mind.

One key to the projected employment of these occupational data items (tasks) is the "signature block" of each task recorded during job/task analysis; another is the computer programming that permits grouping and regrouping of recorded tasks into arrays and hierarchies reflective of representative equipment items, levels of work sophistication, established or innovative occupational structures, or internal task-descriptive hierarchies. To proceed successfully through Training Task Analysis, an important phase of ISD, it is first necessary to provide job task inventories that indicate relationships among tasks, as well as merely list them, and that describe, classify, and catalogue. Such inventories must also be capable of rearrangement of tasks to meet specific requirements by means of a variety of retrieval strategies. These inventories can be built in and by the computer, task interrelationships can be determined, commonality or uniqueness established and measured, and degree of componency and index of complexity fixed. The initial data input is an inventory, to be sure; but, except to serve as a master index of tasks ascribed to a rating, it is not the single or principal such instrument employed in Training Task Analysis in ISD.

This paper will treat a range of inventories and the methodology used to produce and modify them by describing a model developed and currently in experimental use by the Career Development Group, a unit of the Naval Education and Training Program Development Center, Pensacola, Florida. The model shown represents one attempt to secure a massive occupational data input and then to trim it down to an easily manageable catalogue from which to select items for the follow-on steps of ISD.

The task inventory, fully explored and exploited, is more than a list of tasks covering work done within a rating; although the Master Index (figure 1, sample page) is just that. Inventories for use in Front End Analysis (FEA) for ISD meet training task analysis requirements other than those of indexing. For example, inventories can be printed out by equipment hierarchies (platform/system, equipment item, component, module; figure 2), or by established "skill levels" (paygrade groupings), or in divisions or sections specialized to meet other expressed needs of the ISD process. It is the retrieval strategy applied to a multilayered, detailed, and comprehensive occupational data input that make the varied outputs (inventories mentioned above) capable of practical employment in such further steps as Training Task Analysis (TTA).

Principal objectives for the data input design are that the data be detailed, extensive, and reflective less of a technician's opinion than of his recognition and recall of characteristics descriptive of tasks. To this end, tasks to be recorded in FEA are fitted loosely into a data structure that becomes progressively more finite at each lower level of description. This data structure is a Navy world-of-work frame of descending categories of tasks in what eventually becomes an inventory. The major, or first, divisions (Major Functional Categories) indicate the broadest clearly distinct categories of work that tasks fall into, irrespective of the official boundaries of a rating under analysis. The second (next lower) are the Duty Subcategories, work-descriptive areas of smaller compass, within which are the Task Descriptive Characteristics and the Skill Areas (see figure 3).

The first two divisions essentially place the tasks to be recorded; they define or re-define areas of task population. The Descriptive Characteristics and Skill Areas provide extensive and varied items descriptive of task actions and behaviors: skill-related, tool-and-equipment-oriented, explicit actions in a graduated format. It takes the recording of many of these characteristics to make up the task "signature block" data input in the computer; but, in the aggregate, they draw the picture of the task (the signature block is the solid block of numbers from zero (0) to three (3) below the statement "Task Data Worksheet Information" in figure 1).

The initial recording instrument is the Job Data Worksheet (JDW) (figure 5). All job/task information, with the exception of the task signature block and the complexity index is transferred to the computer from this form (compare figures 1 and 5).

The second date recording instrument is the Task Data Worksheet (TDW) (figure 6). This instrument records the appropriate Task Descriptive Characteristics and the applicable Skill Areas, all of which are transferred to the computer, reappearing in the printout as the signature block. It is at this point that the computer actually performs computation. All items appearing in the printout (figure 1)

292

except the complexity index (in figure 1, "Complexity 2.17") represent merely a printed-out restatement of task data input or identification. The complexity index is a computed factor with a range of zero (0) to five (5) resulting from computer manipulation of predetermined weights for the above-mentioned entries in the task signature block.

Meticulous recording of these data provides the Master Index, or total task inventory, for a rating (or any other identified occupational group: NEC, MOS, etc.). The nature and shape of other inventories results from application of a variety of retrieval strategies to the Master Index. Such inventories as that shown in figure 2 (equipment hierarchy) are essentially "shorthand" types of specialized inventories the equipment "level," system-to-module). To get all the data recorded on any specific task, an analyst has only to track the task number back from the specialized (shorthand) inventory to the Master Index.

With no more sophisticated input than that shown, the computer can screen out all identical tasks (all items in a task signature block identical to all such items in other tasks). Task "similarity" (agreed-upon percentage of "identity"), and therefore, "commonality," may be determined by the computer. Using the derived complexity indices and a program designed to explore task interrelationships, "componency" (the degree to which one task is included in, therefore, "component to" another of established higher complexity) may also be determined (see figure 7). Commonality, complexity, and componency of tasks in any inventory are determined by the computer, not by the subject matter personnel who record the data. In the model shown here, the ability of the computer to "look at all tasks with the same eye" (for cataloguing tasks) and to do it reliably and with tireless repetition is fully exploited. Summary decisions, formerly the province of the subject matter expert (SME), have been literally disassembled into numerous and specific items of descriptive data for selection and application to task data recording by the SMEs, recorded by them, then reassembled by the computer into such decision patterns as those mentioned above (complexity, etc.). Few SMEs can match the computer's ability to compare a master inventory of 3,000 to 5,000 tasks for commonality in a single sweep (in some duty subcategories of ratings, sixty percent of tasks recorded proved to be identical, thereby substantially reducing the size of the inventory while not affecting its compass). The philosophy followed in developing this model is that, unless data gathering, opinion infiltration, and actual analysis are recognized as separate (although progressive) steps in task analysis, and this separateness is maintained, opinion infiltration eventually advances in both directions, muddying the entire effort. It is difficult for an SME to compare task #245 with #165, having previously made judgements on the comparative commonality of #102 and #86 without succumbing to the halo effect or some other flowering bias, fatigue, or nagging second thoughts. The computer's monolithic programming leaves it undisturbed by these problems. SME opinion is employed copiously where individual technical expertise, recall of detail, understanding of systems, broad summary judgements, and examination and verification

293

of the computer-made decisions can refine and validate findings and the products of analysis.

```
NO.****110  JOB DATA WORK SHEET INFORMATION               AE-0001-118

RATING  TASK  PACKAGE  TASK-ACTION-CODE  DUTY-SUBCATEGORY
AE      0118  0001     IFT               01
ACTION = IFT (ISOLATE FAULT/TROUBLESHOOT)  ISOLATE FAULT/TROUBLESHOOT WHEEL
                                           WELL LIGHTS 44125

PLATFORM  = P3A/B                                  SYSTEM    = LIGHTING
EQUIPMENT = EXTERIOR LIGHTING                      COMPONENT = WHEEL WELL LIGHTING

CUES, REFERENCES, STANDARDS, ETC., REFERRED TO BY THIS TASK......
CUE.............MALFUNCTION
CUE.............OPERATIONAL CHECK
STANDARD........IAW REFERENCE PUBLICATION
REFERENCE.......NA-01-75-PAA-1-12
REFERENCE.......NA-01-1A-505
TOOLS...........COMMON HAND TOOLS
SUPPORT EQUIP...POWER UNIT NC12/12A
SUPPORT EQUIP...AIR CONDITIONER NB-3A
TEST EQUIP......MULTIMETER PSM-4

TASK DATA WORK SHEET INFORMATION........        COMPLEXITY 2.17
GENERAL       311110000000000000000000000000
DUTY SUB 01   333333300000000000000000000000
SKILL    1    211330000000000000000000000000
SKILL    2    121100000000000000000000000000
SKILL    4    233323200000000000000000000000
SKILL    5    223321113000000000000000000000
```

FIGURE 1   SAMPLE PAGE, MASTER INDEX PRINTOUT

```
1430          ACTION = REMOVE & REPLACE MOD 7352904
PLATFORM      = P-3                                  SYSTEM     = BOMB NAV          AE-0010-0017-02-RAR-2
EQUIPMENT     = AN/ASN-42                            COMPONENT  = NOT CODED
MODULE        = NOT CODED                            COMPLEXITY = 0.65

1431          ACTION = ALIGN  AN/ASN-42 NAV CPTR SET 73520
PLATFORM      = P-3                                  SYSTEM     = BOMB NAV          AE-010-0023-02-ALI-2
EQUIPMENT     = AN/ASN-42 NAV CPTR SET               COMPONENT  = NOT CODED
MODULE        = NOT CODED                            COMPLEXITY = 2.31

803           ACTION = ALIGN CP-632/ASN-42 NAV CPTR 7352400
PLATFORM      = P-3                                  SYSTEM     = BOMB NAV          AE-0010-0320-02-ALI-3
EQUIPMENT     = AN/ASN-42                            COMPONENT  = CP-632
MODULE        = NOT CODED                            COMPLEXITY = 1.61

804           ACTION = REMOVE & REPLACE 4A32 PWR AMP ASSY 7352440
PLATFORM      = P-3                                  SYSTEM     = BOMB NAV          AE-010-0319-02-RAR-3
EQUIPMENT     = AN/ASN-42                            COMPONENT  = CP-632
MODULE        = NOT CODED                            COMPLEXITY = 0.51
```

FIGURE 2 SAMPLE PAGE, EQUIPMENT HIERARCHY PRINTOUT

MAJOR FUNCTIONAL CATEGORY

- * MAINTENANCE
- ** FABRICATION/PRODUCTION
- OPERATIONS
- PERSONNEL SERVICES
- ADMINISTRATIVE SERVICES
- INFORMATION SERVICES (MEDIA)
- MILITARY

DUTY SUBCATEGORY

*A. CHECKING/TESTING/INSPECTING
B. REPLACING/RESTORING ITEMS
C. ADJUSTING/ALIGNING/CALIBRATING
D. REPLENISHING/LUBRICATING
E. CLEANING/PRESERVING

**A. CHECKING/TESTING/INSPECTING
B. DESIGNING/PLANNING/LAYING-OUT
C. CONSTRUCTING/ASSEMBLING
D. EVALUATING/EARTHMOVING
E. DESTRUCTING/DISMANTLING
F. FINISHING/TRIMMING/DECORATING

FIGURE 3 DATA STRUCTURE, MAJOR CATEGORIES

GENERAL

ACCESSIBILITY

(Concerns getting to the object to be worked on)

a. Easily accessible; of little consequence in complexity of task.

b. Moderately accessible, e.g., requires opening drawers, removal of plates, panels, boots, covers, or minor components; climbing, etc.

c. Difficult to gain access, e.g., requires disassembly or removal of other components.

MAJOR FUNCTIONAL CATEGORY: FABRICATION/PRODUCTION

DUTY SUBCATEGORY: DESIGNING/PLANNING/LAYING OUT

SPECIFICATIONS AND MEASUREMENTS

a. Specifications provided; only static measurements required.

b. Specifications provided; dynamic measurements required.

c. Specifications must be derived; dynamic measurements required.

MAJOR FUNCTIONAL CATEGORY: MAINTENANCE

DUTY SUBCATEGORY: REPLACING/RESTORING ITEMS

REMOVAL/REPLACEMENT

a. Simple change of location--requires no fastening/unfastening, e.g., lift, push aside, etc.

b. Dual action--requires fastening/connecting/unfastening/disconnecting in addition to change of locations.

c. Multiple action--requires other supporting actions in addition to fastening/connecting/unfastening/disconnecting and change of location.

SKILL AREA: (5) USING TEST EQUIPMENT

OPERATION

a. Built into system or requires no connection to system and provides automated readings after initial set up.

b. (1) Built into system or requires no connections to system but requires manual step-by-step procedure to obtain readings, or

   (2) Must be connected to system but provides automated readings.

c. Must be connected to system and requires manual step-by-step procedure to obtain readings.

FIGURE 4 DATA STRUCTURE, CONTINUED

RATING HT NEC ____ JOB COMMUNITY DAMAGE CONTROL

TASK OBJECT HIERARCHY — WUC/EIC

I CVA Forrestal Class ____
II HULL ____
III CLOSURES + FITTINGS ____
IV WATERTIGHT DOORS ____
V ____ ____
VI ____ ____
VII ____ ____

MFC	DSC	TASK ACTION	TASK ACT. CODE	CUES	STANDARDS	CONDITIONS/SKILLS						WATERTIGHT DOORS	DOGGING MECH	GASKET						
						REFERENCES	TOOLS	SUPPORT MATERIALS	SUPPORT EQUIPMENT	TEST EQUIPMENT	OTHER CONDITIONS									
1	A	INSPECT	INS	1 2 3	1	1		1,5	1			1								
1	C	ADJUST	ADJ	1	1	1	3-8	1-9	4,5			2								
1	E	PRESERVE	PRS	1	1	1	1,2	1,2,3, 4,9	2,3, 5 6			3								
1	B	REMOVE + REPLACE	RAR	1	1	1	3-8	1	4-6				4							
1	B	REMOVE + REPLACE	RAR	1	1	1	5							5						

JOB DATA WORKSHEET NETPD 1500/4
(Rev. 6-78)

ORIG	OPS	XDGT	QA	COPY	DATA SYS (ORIG)	DP	DATA SYS	FILE	PAGE ____ OF ____

FIGURE 5 JOB DATA WORKSHEET

RATING HT

CUES

1. PMS, quarterly
2. When damaged
3. When ship sustains damage

REFERENCES

1. A-608/m17A-N
2. NAVSEA 0901-LP-920-003
   Chap. 9920

SUPPORT MATERIAL

1. Clean rags
2. Paint remover
3. 1" paint brush
4. Hardwood block
5. Carpenter's chalk
6. Sumbol 2190 TEP oil
7. MIL G-23549 Grease
8. ¼ x 1¼ stick packing
   MIL P-17578 Symbol 1425
   Type I or MIL-G-6032, Type II
9. 3206 oil aluminum oxide
   abrasive cloth
10. Gasket material

TEST EQUIPMENT

1. Tape measure

STANDARDS

1. IAW reference

TOOLS

1. Scraper
2. Wire brush
3. Allen wrench set
4. 12' adjustable wrench (2)
5. 8" normal-duty screwdriver
6. Ball peen hammer
7. Hand chisel
8. Drift punch
9. Welding hood
10. Cutting goggles
11. Electrode holder
12. Chipping hammer
13. Gasket punch

SUPPORT EQUIPMENT

1. Flash light
2. oil can
3. Grease gun
4. Tool box
5. Eye protection shield
6. Bucket
7. Oxy fuel cutting torch
8. Welding machine
9. Electric grinder
10. Soap stone
11. Welding electrodes
12. Spark igniter
13. Leather gloves

OTHER CONDITIONS

JOB DATA WORKSHEET RESOURCES	ORIG	OPS	QA	COPY	DATA SYS (ORIG)	DP	FILE	PAGE ____ OF ____ OF PACKA

FIGURE 5 (CONTINUED) JOB DATA WORKSHEET

TASK DATA WORKSHEET

RATING HTC PACKAGE 01

TASK #	GENERAL SUBCATEGORY	DUTY SUBCATEGORY ___	SKILL AREA 1	SKILL AREA 2	SKILL AREA 3	SKILL AREA 4	SKILL AREA 5
1	1 [✓ a] b c	1 [✓ a] b c	1 a b [✓ c]	1 [✓ a] b c	1 a [✓ b] c	1 a b c	1 a b c
10	2 [✓ a] b c	2 [✓ a] b c	2 [✓ a] b c	2 a [✓ b] c		2 a b c	2 a b c
13	3 [✓ a] b c	3 a [✓ b] c	3 [✓ a] b c	3 a [✓ b] c		3 a b c	3 a b c
26	4 [✓ a] b c	4 a b [✓ c]	4 [✓ a] b c	4 a [✓ b] c		4 a b c	4 a b c
	5 a [✓ b] c	5 a [✓ b] c	5 [✓ a] b c			5 a b c	5 a b c
	6 a [✓ b] c	6 a b c				6 a b c	6 a b c
	7 [✓ a] b c	7 a [✓ b] c				7 a b c	7 a b c
	8 a [✓ b] c	8 a [✓ b] c				8 a b c	8 a b c
	9 a b [✓ c]	9 a b c				9 a b c	9 a b c
	10 [✓ a] b c	10 a b c				10 a b c	10 a b c
	11 a b c	11 a b c				11 a b c	
		12 a b c					
		13 a b c					
		14 a b c					
		15 a b c					

TASK #	GENERAL SUBCATEGORY	DUTY SUBCATEGORY ___	SKILL AREA 1	SKILL AREA 2	SKILL AREA 3	SKILL AREA 4	SKILL AREA 5
77	1 [✓ a] b c	1 [✓ a] b c	1 a b [✓ c]	1 [✓ a] b c	1 a b [✓ c]	1 a b c	1 a b c
80	2 [✓ a] b c	2 [✓ a] b c	2 [✓ a] b c	2 a b [✓ c]		2 a b c	2 a b c
88	3 [✓ a] b c	3 a [✓ b] c	3 [✓ a] b c	3 [✓ a] b c		3 a b c	3 a b c
	4 [✓ a] b c	4 a b [✓ c]	4 [✓ a] b c	4 a [✓ b] c		4 a b c	4 a b c
	5 a [✓ b] c	5 a [✓ b] c	5 [✓ a] b c			5 a b c	5 a b c
	6 a [✓ b] c	6 a b c				6 a b c	6 a b c
	7 [✓ a] b c	7 a [✓ b] c				7 a b c	7 a b c
	8 a [✓ b] c	8 a [✓ b] c				8 a b c	8 a b c
	9 a b [✓ c]	9 a b c				9 a b c	9 a b c
	10 [✓ a] b c	10 a b c				10 a b c	10 a b c
	11 a b c	11 a b c				11 a b c	
		12 a b c					
		13 a b c					
		14 a b c					
		15 a b c					

NAVEDTRAPRODEVCEN 1500/8 (REV 11-77) Page 35 of 61

FIGURE 6 TASK DATA WORKSHEET

```
 #  TAC  RTG   PKG  DS    TK#  ACTION
 1   1  IFT  AE  0004  01  0029  ISOLATE FAULT/TROUBLESHOOT NAV FWD INT CONN BX 429FC
                *TYPE =    01    COMPLEXITY =  1.39
 2   2  IFT  AE  0015  01  0092  TROUBLESHOOT PWR SPLY MDL A1(56259171)
                *TYPE =    01    COMPLEXITY =  1.33
 3   3  IFT  AE  0013  01  0050  ISOLATE FAULT & TROUBLE SHOOT 56950 STANDBY ATTITUDE INDR
                *TYPE =    01    COMPLEXITY =  1.32
 4   4  IFT  AE  0010  01  0332  ISOLATE FAULT/TROUBLESHOOT 4A2 SJM AMP 7352470
                *TYPE =    01    COMPLEXITY =  1.31
 5   5  PCR  AE  0013  01  0030  PERFORM CONTINUITY CHECK 7311700 AN-1931/AJB-3 AMP PWR SPLY
                *TYPE =    01    COMPLEXITY =  1.30
 6   6  IFT  AE  0005  01  0034  ISOLATE FAULT/TROUBLESHOOT ROTOR FLD & HOUS 74719.
                *TYPE =    01    COMPLEXITY =  1.29
 7   7  IFT  AE  0019  01  0086  ISOLATE FAULT/TROUBLESHOOT FLT DIR CPTR 7312100
                *TYPE =    01    COMPLEXITY =  1.27
 8   8  PCR  AE  0013  01  0042  CONTINUITY CHECK 7498900 WPN CONT PNL
                *TYPE =    01    COMPLEXITY =  1.25
DUPLICATE TASK DATA WORK SHEET FOLLOWS *****
 9   8  PCR  AE  0013  01  0039  CONTINUITY CHECK 74988 LABS CONT PNL
                *TYPE =    01    COMPLEXITY =  1.25
10   9  IFT  AE  0004  01  0101  ISOLATE FAULT/TROUBLESHOOT AIR DATA CPTR SET 56460
                *TYPE =    01    COMPLEXITY =  1.11
11  10  IFT  AE  0025  01  0133  TROUBLESHOOT AUTO PILOT ENGAGING CONT C-3385/ASW-1
                *TYPE =    01    COMPLEXITY =  1.11
12  11  IFT  AE  0006  01  0073  ISOLATE FAULT/TROUBLESHOOT INV ASN-42 42212
                *TYPE =    01    COMPLEXITY =  1.10
13  12  IFT  AE  0024  01  0646  ISOLATE FAULT/TROUBLESHOOT MT 3489 57578
                *TYPE =    01    COMPLEXITY =  1.02
14  13  IFT  AE  0003  01  0098  ISOLATE FAULT/TROUBLESHOOT TEMP CONT BOX 41335
                *TYPE =    01    COMPLEXITY =  0.60

        1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18
-----------------------------------------------------------------------------------------------------------------
   1    0    49    39    66    67    72    63    74    72    48    50    58    45
   2   51     0    67    62    53    54    37    52    48    62    58    41    39
   3   41    47     0    71    47    47    36    50    74    47    46    31    37
   4   70    62    71     0    60    54    44    51    65    54    74    41    43
   5   71    54    47    61     0    66    65    71    71    42    53    52    40
   6   78    55    48    55    67     0    68    75    73    37    49    61    38
   7   69    38    37    46    66    69     0    65    72    39    48    47    26
   8   92    55    53    54    74    78    66     0    86    54    45    66    41
   9   62    40    60    54    57    58    57    67     0    36    39    47    28
  10   60    74    55    65    49    43    45    60    52     0    63    39    36
  11   76    71    55    89    63    57    55    51    56    63     0    46    41
  12   80    53    40    53    66    77    59    81    74    43    49     0    28
  13   92    85    81    95    97    80    55    86    74    65    74    49     0
DUPS    0     0     0     0     0     0     0     1     0     0     0     0     0

TOTAL TASKS STUDIED =       14
TOTAL DUPLICATE TASKS =      1
```

FIGURE 7   COMPONENCY COMPARISON

SYSTEMATIC INSTRUCTIONAL VALIDATION THROUGH TESTING

By

Marjorie A. Kuenz, Ph.D.
Fred C. Roberts

Presented at the
1978 Conference
of the
Military Testing Association
Oklahoma City, Oklahoma

October 1978

SYSTEMATIC INSTRUCTIONAL VALIDATION THROUGH TESTING

Introduction

Systematic development, implementation, and evaluation of instruction has gained increasing attention as the aspects of accountability, efficiency, and effectiveness of education or training have received more emphasis. Instructional systems development (ISD) is essentially application of a systems approach to educational process. Steps in this approach are basically (1) determining instructional needs, (2) developing effective and efficient solutions to these needs, (3) implementing solutions, and (4) assessing the degree to which these needs are met.

For new instructional programs, ISD can be logically and effectively applied. However, for existing programs, a comprehensive testing plan will provide an effective alternative. The testing plan is designed so that, if ISD is later applied to the instruction, the methods used and data collected will be applicable to and consistent with ISD. The purpose of this paper is to present a technique for validating instructional programs through course testing instruments in order to supplement the development process used.

Testing integral to instruction

Testing serves two main purposes in Navy health sciences education and training: (1) to assess student knowledges and skills acquired while participating in training activities; and (2) to assess carryover of knowledges and skills to real-life/actual job settings. For these purposes, at least three aspects must be measured: cognitions, motor skills, and application of cognitions and skills in the job setting. For each of these aspects, numerous instructional objectives exist for a given course, these objectives giving specific substance to this otherwise theoretical distinction. Test items are designed to represent and conform to objectives and the methods of instruction.

For testing to assess the effect/success of instruction, it is essential that tests measure the outcome of instruction at whatever level of detail the instruction is given. The key to determining the effectiveness of instruction is the precision with which what is taught is tested. Test items must measure specific behavior, with the conditions under which the behavior is to be achieved and the manner in which the behavior and conditions are to be demonstrated established by the objectives.

The number of written test items and performance assessments that can be generated to adequately represent all the instruction conveyed in a particular course or program is almost always more than can practically be administered to any student. Sampling of instructional content or of testing mechanisms is usually done to reduce the amount of actual testing to a proportion of the total. Selection of test items and instruments for use at any one time can be done by random or stratified sampling procedures or a combination of both, the objectives of the instruction will usually guide the choice of sampling procedures. Whatever the selection process used, all testing mechanisms need to be validated beyond the face and content validity built in during development. Typically, this validation takes the form of concurrent or predictive validity studies where appropriace criterion measures are available or developed, against which the new tests are compared.

Field validation

A different approach to validation, however, is more appropriate for specialized training particularly when ISD has not been used in program development. This approach is closely linked to the second purpose of testing: to assess carryover of knowledges and skills to the real-life/actual job setting. Validation through testing can be accomplished only if direct input is obtained from appropriate "field" specialists or practitioners. In traditional curricular development, tests are devised to correspond to instruction. It is essential, however, to extend validation by determining the extent to which instructional content and tests correspond to job requirements.

The process by which tests can be validated against job requirements can be applied to any type of testing mechanism. Two specific examples will be given here, one of written test items and one a performance rating scale which are part of a 16-week (640-hour) course for otolaryngology (ENT) technicians. This course consists of five content units: anatomy and physiology, ENT surgery, clinic technique, operating room procedures, and audiology. The expressed purpose of the course is to "provide trained enlisted personnel with the knowledges and skills needed to assist medical officers in the treatment and care of patients with otolaryngology disorders" (Catalog of Navy Training Courses, June, 1978, Vol. 2).

Validation of written test items

For field validation of test items for this course, a sample of Otolaryngologists (ear, nose, and throat (ENT) specialists) was chosen based on the following criteria: [illegible] the physicians were on a hospital staff (Navy Regional Medical Center); (2) three or more ENT technicians were assigned to assist the physicians in the clinic and operating rooms of the hospital; and (3) the physicians were the immediate supervisors of one or more ENT technicians.

The physicians were directed to judge how important information contained in each test item was for the technician's performance of his clinic and operating room duties. To obtain these judgements in a systematic way, test items were presented in a rating scale format--each item preceded by five response columns.

The statement to which the physician responded was: "The item tests information that is essential to the technician's job performance."

Judgement of the importance of the content of the test question was expressed by indicating how much he agrees or disagrees with this statement. The five columns located to the left of each item were labeled as: SA (Strongly Agree), A (Agree), U (Undecided), D (Disagree), and SD (Strongly Disagree). The specialist was instructed to mark an "X" in the appropriate column to represent his judgement about the essentialness of the technician knowing the item's content. (See Tables 1.1 and 1.2). At the end of each section of the test--there were five sections corresponding to the five content units of the course--was a Comments page on which the physician could note topics that were not included but which should be tested.

Responses to the rating scale were received from 8 of the 9 Regional Medical Centers. A frequency tally was done of the ratings given each test item. An item for which more than half of the ratings fell below the midpoint, i.e., five or more responses were in the columns of "Disagree" and "Strongly Disagree," was considered to be judged non-essential. Of the original 200 test items submitted for review by the Otolaryngologists, 45 were judged to test non-essential information.

The remaining 155 items were then revised as recommended by the reviewers and submitted to the ENT technician instructional staff for review, revision, and additions. Instructors were requested to perform two functions: (1) verify that the remaining test items correspond to instructional objectives, or revise one or the other so that they do correspond, and (2) propose additional test items to measure objectives not represented by the remaining test items, indicating the objective being measured. The revised and new test items are then submitted for field validation in the manner described above, the process being an iterative one.

Validation of performance rating scale

Since a large portion of ENT technician instruction consists of skill development, assessing the level at which these skills are performed in the real-life/actual job setting is the most appropriate method to determine the adequacy of instruction for these skills. Field validation of the skills for which ENT technicians were trained was initiated through a training follow-up or feedback instrument. Because of the diversity of tasks for which the technician is trained--this primarily due to his assisting in either or both clinic and operating room setting--two forms were developed: one related to clinic tasks and one for tasks performed in the operating room. Initial review and refinement of task statements was accomplished in cooperation with an ENT physician and a senior ENT technician.

Within 30 days of completion of instruction, training follow-up forms were sent to the duty station of the graduates of the school. Otolaryngologists supervising the recent graduates were requested in a cover letter to complete the forms for purposes of assisting "in determining the relevance of the Otolaryngology Technician training curriculum."

Each of the two training follow-up forms consisted of a list of tasks for which the technician is trained, each task followed by three response columns. Specific instructions to the physician completing the form included:

> Attached is a list of tasks an ENT technician may be required to do in the clinic (or operating room). If the specified technician is assigned for the equivalent of one day or more per week to the clinic (or operating room), this Clinic (or Operating Room) Assignment form should be completed for him/her. . . .
>
> In the Columns numbered I, II, and III following each task, indicate specific information about the technician's performance of that task.
>
> Column I: "Does the technician perform the task?" Mark an "X" under either "YES" or "NO", whichever is appropriate.
>
> Column II: Use this column only if the technician performs the task (if you marked an "X" under "YES" in Column I).
>
> "How well does the technician do the task?" Mark an "X" in the block under the term that best describes the quality of this performance, namely, EXCELLENTLY, ADEQUATELY, or INADEQUATELY.
>
> Column III: Use this column only if the technician does not perform the task (if you marked an "X" under "NO" in Column I).
>
> "What is the reason that the technician does not perform the task?" Mark an "X" to indicate which of the following reasons is appropriate:
>
> 1. The technician says he/she wasn't taught how to do it.
> 2. The technician doesn't know how to do it.
> 3. Operating room procedures, or your way of practice, does not require the technician to do the task.

Allowing for incompleteness in the list of tasks, the specialist was also requested to supply in the space provided under Additional Duties, those tasks that the technician does which were not included in the list. A Comments section was also provided, with the specific request that the physician give general suggestions he may have regarding the follow-up itself. Depending on the actual work assignments of the technician, either the clinic or operating room form (or both) was/were completed for each graduate.

Summary of performance data

The initial group of recent graduates whose performance was assessed consisted of 14 ENT technicians. For this group, responses for 12 were received. A frequency tally was done of the responses to the three questions asked: if the technician performs the task, how well he performs, or why he doesn't perform it. Totaling the responses initially provided the following data:

(1) the number of recently graduated ENT technicians who perform and do not perform each listed task;

(2) the number who are judged to perform each task at the three specified levels of competence; and

(3) the number of technicians who do not perform the tasks for each of various reasons.

Summary descriptive statistics were then calculated for each task, providing the following further data:

(1) the proportion (or percentage) who perform and do not perform each task;

(2) the average (median) competence rating given for those who perform each task (for purposes of calculation, a rating of Excellent was converted to 3, Adequately to 2, and Inadequately to 1); and

(3) an index of variability in ratings (semi-interquartile range) using the same numerical conversions.

Additional tasks supplied by the physicians were summarized in the same way.

Application of data

The actual number of responses from otolaryngologists to the field survey of the essentialness of test item content and the performance of recent graduates is insufficient to warrant extensive curricular revision. The process, however, is being repeated for additional tests and subsequent graduating classes to substantiate trends and to clarify topics and tasks for which responses varied greatly. The manner in which these types of data can be utilized for validation and revision of tests and instruction are straightforward, however.

Test validation and revision

1. Written test items judged to contain non-essential information are eliminated from the usable item pool. 2. Test items that are judged to contain inaccuracies, based on comments of specialty practitioners are

309

revised accordingly and validated, as are new items. 3. Similarly, tasks performed by those completing instruction constitute the list of tasks for which others should be trained, and, therefore, performance of those tasks is what is testable. 4. A pool of field validated items and tasks is maintained from which tests for specific purposes and according to specified parameters can be drawn. 5. Periodic re-validation of testing instruments will be implemented so that changes in knowledge and technology can be represented and incorporated.

Instructional validation and revision

1. Those areas of content judged essential and the skills reported as functional by field practitioners form the basis for instruction.
2. That content judged non-essential and tasks not performed are removed from instruction (unless emergency or contingency consideration require its being retained).
3. Recommendations for additions or deletions to instruction or testing are compared with data from field practitioners. If validation data is not available, it is collected using one of the previously described procedures.

Conclusions

While procedures for revising instruction and testing are often organizationally specific and tied to considerations not at all a part of educational process, the need to firmly base such revisions on the real-world considerations is almost too obvious to mention. The all-too-common and cyclic process of instructor determining what should be instructed, most often with real sincerity, believing he is the best judge of what should be taught because he has been teaching it for *N* years, needs to become instead an interactive process. Obtaining and incorporating "field" data into instruction and instructional development is essential and efficient if your goal is validity.

AUDIOLOGY

The item tests information essential to job performance.					
SA	A	U	D	SD	
					1. What frequency span is used for the short increment sensitivity index? a. 4K, 2K, 1K, 500Hz, 250Hz b. 6K, 4K, 3K, 500Hz, 250Hz c. 6K, 4K, 3K, 2K, 500Hz, 250Hz d. 6K, 4K, 2K, 1K, 500Hz, 250Hz
					2. Nonsyllabic, phonetically balanced, and equally difficult words are characteristics of the ____________ test. a. short increment sensitivity index b. Stenger c. speech reception threshold d. speech discrimination
					3. What is the most efficient type of masking noise for pure tones? a. speech b. sawtooth c. white d. narrow band

Table 1.1

AUDIOLOGY

The item tests information essential to job performance.

For items 1-4 select from column B the term which best fits the definition in Column A.

Column B

a. air conduction
b. sensorineural hearing loss
c. bone conduction
d. audiometer
e. conductive hearing loss

SA	A	U	D	SD	Column A
					1. d a device designed to determine the quantity of hearing
					2. a transmission of sound stimuli to the eardrum via the external ear canal
					3. b hearing loss caused by decreased sensitivity of the end organ of hearing
					4. c transmission of sound vibrations to the inner ear via the bones of the skull
					5. What examination determines the ability of a patient to understand what he hears? [✓ a. speech discrimination] b. speech reception threshold c. short increment sensitivity index d. Stenger test

Table 1.2

OTOLARYNGOLOGY (ENT) TECHNICIAN (8446) TRAINING FOLLOW-UP:

O.R. ASSIGNMENT

ENT Technician ____________________

Station ____________________________

Rater ______________________________

Technician has been working in clinic/O.R. for ____ months.

TASKS,

TRANSPORTING THE SURGICAL PATIENT TO THE O.R.

	I		II			III		
	Performs task		If Yes, how well			If No, give reason		
	Yes	No	Excellently	Adequately	Inadequately	Was not taught	Does not know how	Not Required
Verify identification of patient using arm band and chart.								
Review pre-operative checklist for completeness.								
Transfer patient from bed to guerney/crib.								
Push occupied guerney/crib to O.R.								
Transport patient on Stryker or Foster frame.								
Transport patient requiring oxygen, IV, or special care.								
Transport patient in orthopedic traction.								
Observe patient for signs of chilling.								
Watch for and report symptoms of external hemorrhage.								
Watch for and report abnormal respirations.								

Table 2.1

REFERENCES

1. Catalog of Navy Training Courses (CANTRAC). Vol.2. June, 1978.
2. Chief of Naval Training, Instruction on Measurement of Student Achievement. August, 1973.
3. Dick, Walter and Carey, Lou. The Systematic Design of Instruction. Glenview, Ill.: Scott, Foresman & Co., 1978.
4. Interservice Procedures for Instructional Systems Development, (NAVEDTRA 106A). Vols.1-4. August, 1976.
5. Morreau, L. E.. "The Structural Analysis and Classification of Objectives," Educational Technology. March, 1974, 46-48.
6. Popham, W. J., and Shapiro, D. "Measurement Requisites for Competency Assurance in the Health Profession," Evaluation and the Health Professions. Vol.1., No.1. Spring, 1978.

SCHEDULING FORMAL SCHOOL TRAINING TO MAXIMIZE COST EFFECTIVENESS

DOUG GOODGAME
OCCUPATIONAL RESEARCH PROGRAM
TEXAS A&M UNIVERSITY

ABSTRACT

Procedures for designing instructional systems which rely upon the job inventory method to collect occupational data from incumbent workers and job supervisors, can, in the data analysis phase, provide the designer information for making decisions on cost effective scheduling of formal school training. Two situations are presented to substantiate this assertion. One situation describes correlational relationships between task factor ratings (measuring work requirements at the job site) which dictate that formal school training should be scheduled prior to job assignment. The second situation reveals relationships whereby formal school training may be delayed indefinitely.

The results of three occupational studies are reviewed to demonstrate sample applications. The studies reveal that uniform relationships do not exist across work requirements in similar occupations and indicate that unique conditions in the work environment affect relationships between work requirements.

317

I INTRODUCTION AND STATEMENT OF PROBLEM

The purpose of this article is to demonstrate a method whereby instructional system designers can determine if formal school training should be scheduled prior to job assignment. In the event formal school training can be delayed to a period after job assignment the designers will then be able to develop less costly forms of training and implement the training during an on-job-training phase. Cost of training is often related to the location where training is delivered. These locations can include, but are not necessarily limited to the following:

On The Job: Training experiences are directly keyed to job actions and easy to learn job practices, and procedures. Supervisors and senior workers control the content, pace and sequence of instruction.

Agency Classroom: Training experiences are often keyed to policy, procedures and specialized job functions of the employing organization. Management and staff from the employing agency control content, pace, and sequence of instruction.

Remote Classroom: Training experiences are directed to those work behaviors and technologies most difficult to learn. Training at this location (referred to as formal school) represents a pooling or sharing of training resources where instructional specialists control the content, pace and sequence of instruction.

The most costly training occurs in the formal school setting at the remote classroom. Training costs at this location are the cumulative result of trainees loss of production to attend the school or cost to replace trainee with worker of comparative ability. Additional costs include trainees travel and per diem plus the cost to support instructional resources at the remote location. Many of these costs can be minimized if initial training can be delivered on-the-job or in the agency classroom reserving formal school training to a later more cost convenient period. Appropriate on-job and agency classroom training can also reduce time and cost to administer formal school training by addressing skills and knowledges that are readily learned in those training environments. In addition, work experience at the job site can provide valuable learning experiences and develop a foundation and frame of reference for formal school training.

It is not always possible to delay formal school training to a later, more cost convenient period in a trainees work experience. In many situations work requirements at the job site necessitate the aquisition of critical knowledge and skills before a worker can function productively in the assigned work environment.

Determining if formal school training can be delayed without violating critical work requirements at the job site is the central problem to be addressed by this article. The solution to this problem requires an analysis of occupational data measuring work requirements of tasks performed by incumbent workers at the job site.

II REVIEW OF RELEVANT LITERATURE

Considerable research has been conducted to develop appropriate methodologies for designing instructional systems to solve training problems (1) (8). To date, the effort has concentrated on job-task analysis techniques and procedures for translating results of task analysis into curriculum. These activities mark the beginning points for instructional system design. Designers often assume that the end product will be delivered in the most cost effective manner on a schedule consistent with work requirements at the job site. Too often well designed instructional systems are not delivered on a schedule consistent with work requirements. This is unfortunate, for designers are now beginning to collect the types of occupational data which make such determinations possible.

An investigation to determine if formal school training should be scheduled prior to job assignment can be a by-product of standard procedures for conducting job analytic studies in an occupational area. There is little additional work required of a designer of instruction systems provided the designer follows recommended procedures and collects, for analyses, specific types of occupational data using job or task inventories (2) (7).

III METHODOLOGY

Data Requirements

Numerous organizations presently follow recommended procedures in constructing job or task inventories which enable large samples of incumbent workers in an occupational field to report performance and non-performance of tasks across their job domain. A job or task inventory, if correctly developed, will contain a listing of all tasks performed by incumbent workers in a specific job domain. Each incumbent can then use the task inventory to report the unique set of tasks performed at the job site. Task level job descriptions can be computed for a group of incumbent workers to report the percentage of workers performing each task.

The percentage performing value is a vital measure of emphasis of task performance at the job site and identifies what workers do and do not do as they routinely perform their work assignments. Resultant values produce a data vector across all tasks in the job domain with values ranging from 0% to 100%. This data vector (percentage of members performing tasks) or task factor will be referred to in an abbreviated form as "PERP" in this article.

Job analytic studies, conducted to design instructional systems, also require that certain types of occupational data be collected from experienced job supervisors to define critical work and training requirements of tasks. To accomplish this, job supervisors review each task and report ratings (using specially designed Likert scales) on each of the following task factors:

- -Task learning difficulty (TLD): time required to learn to perform a task satisfactorily. (low scale values equal short learning periods)
- -Task delay tolerance (TDT): delay time tolerated prior to beginning performance of a task once incumbent observes that task must be performed. (low scale values equal short delay times)

-Consequences of performance failure (CPF): severity of consequences of inadequate performance of task. (low scale values equal inconsequential results)

Inter-rater reliability coefficients should be computed on ratings from each factor to identify and delete unreliable raters from the investigation (6). Resultant means provide a measure of the work requirements for each task on each task factor and establishes a data vector for each factor.

Recent studies have demonstrated that data vectors for each of the four task factors presented in this article (PERP, TLD, TDT, and CPF) can account for 80 to 90 percent of the variance in a criterion data vector representing reliable ratings on training priority of tasks (3) (4) (5). It is evident that a task's estimated priority for training is a function of: a) emphasis on task performance at job site, b) task delay tolerance, c) task learning difficulty and d) consequences of task performance failure.

The associative variations among these task factors (factor vectors) can present very intriguing glimpses into the work requirements for an occupation. These variations, in a correlational framework, can allow designers of instructional systems to determine if delay can be tolerated in delivering formal school training. In this regard, the next section presents two examples: the first identifies certain relationships among work requirements that necessitate delivery of formal school training prior to job assignment, and the second identifies an opposite set of relationships indicating that formal school training can be delayed indefinitely.

IV PROCEDURES FOR ANALYSING DATA

The first step in analysing the work requirements of an occupation relative to the four task factors requires computing and reporting a correlation matrix. The matrix reports the Pearson product-moment correlation coefficient between each factor vector and all other factor vectors in an occupational study.

The correlates between factor vectors in an occupational study can reveal to the designer relationships between work requirements at the job site, which, in turn, can help the designer determine whether delay in formal school training would seriously violate work requirements of tasks routinely performed at the job site.

The following is being presented as an example of a situation where formal school training, at the remote location, should be scheduled prior to job assignment. The correlates in a model matrix should indicate that:

1. PERP. is negatively correlated with TDT: This implies that for tasks performed by a majority of workers, the workers have little delay time in initiating performance of the tasks once the workers observe that the task has to be performed. It also implies that workers may not have time to consult a supervisor or senior worker or look up a procedure in a manual before initiating performance of the task.

2. PERP. is positively correlated with TLD: This implies that tasks performed by a majority of workers are difficult to learn to perform. Difficulty being expressed as time required to learn to perform a task satisfactorily.
3. PERP. is positively correlated with CPF: This implies that tasks performed by a majority of the incumbent workers will result in severe consequences if not performed correctly.
4. TDT is negatively correlated with TLD: This correlation indicates that tasks with low time delay tolerances require longer periods of learning time.
5. TDT is negatively correlated with CPF: This implies that tasks with low time delay tolerances will result in severe consequences if performance failure occurs.
6. TLD is positively correlated with CPF: This correlation implies that tasks which are difficult to learn to perform correctly will result in severe consequences in the event of performance failure.

Correlates of high magnitude in the above example would apply to few jobs in our work society. It is highly probable the correlates would apply to tasks performed by emergency medical service personnel and firefighters to name two occupations where the job demands are exceedingly rigorous with task performance constrained by low time delays. It is conceivable that an analysis of occupational data from these two areas would indicate that formal school training should occur prior to job assignment.

A reverse in the signs associated with the correlates between factor vectors presented in the model matrix will establish the boundaries for a second matrix. This second model would indicate a high probability that formal school training could be delayed indefinitely. Such a reverse implies that a majority of the tasks performed by workers will exhibit high task delay tolerance values, be easy to learn to perform, and produce inconsequential results if performance failure occurs. In addition, tasks with low time delay tolerances will be easy to learn to perform and will produce inconsequential results if performance failure occurs. Also, tasks that are difficult to learn to perform will produce results in which performance failures will be inconsequential.

The two correlational models presented in this section represent situations in which occupational work requirements dictate two extremes. The first model implies that formal school training should be scheduled for new employees prior to job assignment, since an analysis of work requirements indicates that a new employee would have difficulty performing assigned tasks without special training. The second model implies that formal school training could be delayed indefinitely, since the analysis of work requirements indicates a high probability that a new employee would not have any difficulty learning to perform assigned tasks at the job site.

In the next section correlates between factor vectors generated from three occupational fields are reviewed to demonstrate field application of the process.

V THREE SAMPLE APPLICATIONS

The Occupational Research Program at Texas A&M University recently conducted job analytic studies in three criminal justice occupations to derive training requirements for designing instructional systems. In one study 295 tasks performed by 258 county detention officers were analyzed (5). Tasks performed in county detention centers are closely related to tasks performed by correctional officers in state and federal correctional institutions. Generally, county detention officers process prisoners into the center, supervise the custody of inmates housed in cell blocks and process prisoners for release from custody.

A second study investigated the work performed by 121 sheriffs' deputies (4). A portion of this study focused upon 423 tasks performed by deputies working in counties with less than 40,000 population. These officers perform a myriad of county law enforcement and public service tasks.

The third study analyzed 355 tasks performed by 47 field sergeants working in police departments serving highly populated cities (3). These officers supervise the work of uniformed patrolmen who provide law enforcement and public assistance services to municipal government.

The table on page 8 reports a matrix of correlates between factor vectors across three occupations. The notation "PERP X TDT" in item 1 below refers to two factor vectors of interest. The notation r_A, r_B, and r_C refers to correlates in each occupational field relative to the factor vector of interest. A review of the findings indicates that:

1. PERP X TDT: ($r_A = -.45$, $r_B = -.50$, & $r_C = -.35$)

   A majority of the officers in each occupation perform tasks where low time delays are tolerated prior to initiating performance of a task once an officer observes that a task has to be performed. This implies that officers may not have time to seek assistance or guidance from supervisors or fellow officers on how to perform a task, nor be able to look up a procedure in a manual.

2. PERP X TLD: ($r_A = -.46$, $r_B = .35$, & $r_C = -.17$)

   A majority of the officers in occupations A and C perform tasks which are relatively easy to learn to perform as indicated by the negative coefficients. This is not the case with deputy sheriffs working in less populated counties. Here, a positive coefficient implies the tasks performed by a majority of the officers are difficult to learn to perform; a reverse of the situation normally expected of workers in an entry level position. It is generally understood that these deputies perform a wide range of tasks which in larger counties would be performed by senior deputies or specialists.

322

3. PERP X CPF: ($r_A = .24$, $r_B = .48$, & $r_C = -.02$*)

A majority of the officers in occupations A and B perform tasks in which the consequences of performance failure was deemed very severe. This is evidently not true for officers in occupation C. The job descriptions for officers in this field, revealed that field sergeants continue to perform many line tasks. Line tasks being the type of work normally performed by uniformed patrolmen.

4. TDT X TLD: ($r_A = -.06$*, $r_B = -.34$, & $r_C = -.20$)

It appears for occupations B and C that a significant negative correlation exists between the length of time required to learn to perform a task and the delay time tolerated to initiate performance at the job site. This was not a characteristic of the relationship between tasks performed by county detention officers as evidenced by the low coefficient $r = -.06$.

5. TDT X CPT: ($r_A = -.77$, $r_B = -.76$, & $r_C = -.59$)

For these occupations a high correlation exists between the delay time tolerated prior to initiating performance of a task and the resultant severity if performance failure occurs. This implies that tasks with low time delay tolerances will produce severe consequences if not performed correctly.

6. TLD X CPF: ($r_A = .39$, $r_B = .71$, & $r_C = .71$)

For these occupations a high correlation exists between the time required to learn to perform a task satisfactorily and severity of consequences if performance failure occurs. Specifically, this indicates that tasks requiring long periods of learning time will, if not performed correctly, produce severe consequences.

*Coefficients were not deemed significant at .05 level.

VI CONCLUSIONS

The correlates report very promounced relationships between work requirements in each occupation, but indicate that uniform relationships do not exist across these occupations. It could have been assumed that all law enforcement and detention related occupations in criminal justice career fields would exhibit similar relationships between work requirements across all occupations.

According to the first correlational model outlined in Section III it would be appropriate for sheriff's deputies to receive formal school training prior to job assignment since the work requirements of performed tasks meet all six critical criteria. It is possible that formal school training could be delayed for new employees in county detention centers since a majority of the officers perform tasks which are not difficult to learn to perform. And, it is conceivable that supervisory training can be delayed for newly appointed first-line supervisors since a majority of supervisors continue to perform line tasks which relative to all tasks in their job domain, are easy to learn to perform. Also, a majority of the supervisors perform tasks where consequences of performance failure ratings (from insignificant to serious) appears randomly distributed across all tasks.

TABLE I

Correlates Between Task Factors Across
Three Criminal Justice Occupations

OCCUPATIONS

A = 258 County detention officers, 295 tasks
B = 121 Deputy sheriffs, 423 tasks
C = 47 Field sergeants (municipal police department), 355 tasks

TASK FACTORS (DATA VECTORS)

PERP = Percentage of members performing tasks
TDT = Task delay tolerance
TLD = Task learning difficulty
CPF = Consequences of performance failure

	TDT			TLD			CPF		
	A	B	C	A	B	C	A	B	C
PERP	-.45	-.50	-.35	-.46	.35	-.17	.24	.48	-.02
TDT	1	1	1	-.06	-.34	-.20	-.77	-.76	-.59
TLD				1	1	1	.39	.71	.71

VII SUMMARY

Designers of instructional systems need to determine if formal school training should be scheduled prior to job assignment. In the event formal school training can be delayed less costly forms of training can often be instituted at the job site. This training can provide job experiences and instruction which will benefit the employee during his formal school experience. The job experience will provide a frame of reference to make formal school training more job related, and instruction at the job site and agency classroom can build knowledge and skills which may permit reduction in amount of time required to deliver formal school training.

Present procedures for designing instructional systems incorporate techniques for collecting data to validate the job relatedness of proposed training curriculum and can define critical tasks which new employees should be trained to perform. This same data, when analyzed in a correlation matrix, can offer a designer of instructional systems insights into the critical work requirements of tasks distributed across a specific job domain.

Determining if formal school training can be delayed requires a special analysis of the critical work requirements at the job site. The analysis involves computing a matrix of correlates among task-factor vectors measuring: a) emphasis of task performance at the job site, b) task learning difficulty, c) task delay tolerance, and d) consequences of performance failure of tasks. The resultant matrix will enable the designer to assess relationships between work requirements and determine if formal school training can be delayed.

Length of delay is a judgment the designer will have to make based on knowledge of when tasks which are difficult to learn to perform and have low time delay tolerances become major assignments for new employees. An advisory committee composed of knowledgeable first line supervisors can assist the designer in setting time limits which can vary according to the work environments at various job site.

REFERENCES

1. Center for Vocational Education, "Performance Content for Job Training", Ohio State University, March, 1977.

2. Christal, R.E., & Weissmuller, J.J., "New CODAP Programs for Analysing Task Factor Information". AFHRL-TR-76-3, AD-A026 121. Lackland AFB, Texas. Occupational and Manpower Research Division, Air Force Human Resources Laboratory, May 1976.

3. Goodgame, D.T., and Hogue, K.C., "Analysis and Definition of Critical Training Requirements for First-line Supervisors in Municipal Police Departments". Occupational Research Program, Industrial Engineering Department, Texas A&M University, June, 1978.

4. Goodgame, D.T. and Hogue, K.C., "Analysis and Definition of Critical Training Requirements for Sheriffs' Deputies". Occupational Research Program, Industrial Engineering Department, Texas A&M University, June, 1978.

5. Goodgame, D.T., and Hogue, K.C., "Validation of Critical Training Requirements for County Detention Officers". Occupational Research Program, Industrial Engineering Department, Texas A&M University, June, 1978.

6. Goody, K., "Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to Identify Divergent Raters". AFHRL-TR-76 82, AD-A034 327. Lackland AFB, Texas, Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October, 1976.

7. Goody, K., "Task Factor Benchmark Scales for Training Priority Analysis: Overview and Developmental Phase for Administrative/General Aptitude Area." AFHRL-TR-76-15, AD-A025 847. Lackland AFB, Texas, Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June, 1976.

8. Morsh, J.E. & Archer, W.B., "Procedural Guide for Conducting Occupational Surveys in the United States Air Force". PRL-TR-67-11. Lackland Air Force Base, Texas. Aerospace Medical Division, Personnel Research Laboratory, Sept., 1967.

METHODS FOR DETERMINING SAFETY TRAINING PRIORITIES FOR JOB TASKS

By
Nancy A. Thompson
and
Hendrick W. Ruck
Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, Texas

The Air Force Human Resources Laboratory is actively involved in performing training requirements research using both operational and experimental occupational survey data. Currently, research is aimed at providing products to assist training designers in deciding which tasks should be considered for training in various Air Force specialties. In addition to the training requirements work, a basic research study is presently being conducted on the feasibility of developing a method of prioritizing job tasks in terms of hazard potential, expected frequency of accidents and other pertinent factors that could assist training designers in determining needs for safety training.

There are several similarities between the objectives of the safety training research and the training requirements research. Both streams of research endeavor to define certain task factors that will prove to be predictive of training requirements. Both efforts employ the regression modeling approach, a method that is more thoroughly discussed by Ruck (1978). Also, both projects share the goal of contributing meaningfully to the job relevancy of Air Force training programs.

The safety training research described in this paper is in response to a request from the Air Force Inspection and Safety Center (AFISC) at Norton Air Force Base. The objective is to provide the AFISC with information to help prevent on-the-job accidents that result in injuries, loss of equipment, loss of time and loss of materials. The approach is to collect hazard potential ratings for technical tasks and determine the extent to which these hazard potential ratings (and a number of other task factor ratings) can predict accidents on the job. The purpose of this paper is to present an approach to working with accident data, to discuss some of the problems associated with this type of data and to discuss future directions for an expanded study of accident data in various career fields. It is important to note that the problem addressed in this paper has to do with "what tasks will have accidents" and not with "which people will have accidents." Therefore, the question addressed in this paper is somewhat different from that normally considered in safety research.

327

Approach

The approach taken in the safety training priority research was to define task factors believed or known to be predictive of accidents, to collect task factor ratings from experts, to prioritize job tasks in terms of need for safety training, and to develop regression models with predictive efficiency for safety training. Data were analyzed using the comprehensive occupational data analysis programs (CODAP) (Christal & Weissmuller, 1976).

Several alternatives to determining characteristics peculiar to safety training were considered. Based on previous training priorities research, ratings of consequences of inadequate performance, task delay tolerance and task difficulty were included in the analysis. Consideration was given to a scale that would yield ratings of safety training requirements, but was rejected because the scale was not clearly related to the problem. Another possible approach was to use only tasks which had been involved with accidents; however, this approach did not address tasks that might have been potentially hazardous but had not had any occurrence of accidents. Ultimately a new task factor scale was devised to measure the hazard potential of tasks. The approach for this initial study is promising in that rater response has been good and initial results are encouraging.

Data Collection

The aircraft armament specialty (AFSC 462X0, previously called weapons mechanic), was chosen for the present study. The aircraft armament career ladder consists of 12,669 incumbents, 2,588 of which serve at a supervisory skill-level. The major job groups for non-supervisory incumbents are weapons loader (72%), weapons release (18%), and gun services (10%). Each job incumbent performs an average of 70 tasks out of a possible 527 tasks included in the job inventory. Twenty-nine percent of the time spent by job incumbents is on supervistory functions; 28% of their time is spent on loading functions; and 15% is spent on flight line inspections and operational checks.

Criterion data were extracted from accident reports that were supplied by the AFISC. Among various variables, the reports provided the accident location and date, the cost per accident, and a narrative describing the accident. These reports were reviewed by a person knowledgeable in the aircraft armament specialty and the accidents were matched with the tasks that were being performed when the accidents occurred. The number of accidents per task was then established for each of 527 tasks as listed in the job inventory for 462X0. As is frequently discovered when dealing with accident data, the ratio of accidents to tasks was very low. In a time frame beginning in July, 1975, and ending in December, 1976, a total of only 49 accidents was reported that could be related to the job inventory for the aircraft armament specialty. Furthermore, these 49 accidents were associated with only 20 tasks from the 527 tasks included in the inventory. When the number of accidents was broken down by duty, it was found that almost half (26) of the accidents occurred while performing loading tasks. Other accident related duties were: (a) performing operational checks of aircraft suspension, release, launch, and monitor and control systems (10 accidents); (b) shipping and transporting munitions (8 accidents); (c) performing flight line inspections of aircraft suspension, release, launch, and monitor and control components (2 accidents); (d) performing flight line maintenance of gun systems (2 accidents); and (e) removing, installing, and replacing aircraft suspension, release launch, and monitor and control components (1 accident).

Several task factors were collected from the supervisors in the field. These factors included consequences of inadequate performance, task delay tolerance, and task difficulty. The development of these factors was described by Mead (1975) at the 17th annual conference of the Military Testing Association.

Since this study was concerned with safety training, a fourth factor was developed called hazard potential. The hazard potential factor was suggested by a study which evaluated human effects on nuclear systems safety (Askren, Campbell, Seifert, Hall, Johnson, Sulzen, 1976). The hazard potential scale was designed to determine which tasks are more hazardous to perform than others and might, therefore, cause accidents. If the raters agree that certain tasks are more hazardous to perform than others, then safety training can be recommended for those tasks.

The nine point hazard potential scale ranges from extremely low hazard potential through extremely high hazard potential. The hazard scale was sent to seven and nine skill level supervisors who were asked to first check only those tasks in the inventory which he or she considered to be potentially hazardous. Then the rater was asked to rate the checked tasks on a scale from 1-9 to indicate how potentially hazardous each task is. For analysis purposes the scale was considered a 10 point scale because a task not checked was given a value of zero, indicating no hazard potential. Appendix A illustrates the rating scale. Appendix B presents the inter-rater agreement indices for a sample size of [illegible] (based on the Spearman Brown formula) for all task factors.

In addition to the four task factors, two other variables which had previously been collected in a routine occupational survey were considered: percent members with 1-48 months total active military service performing each task and an index of percent of time spent by members with 1-48 months service performing each task. Appendix C shows the zero-order correlations among the six variables. Although the correlation between hazard potential and consequences of inadequate performance is high, (r=.70), there are some conceptual differences in the two factors. Other factors correlate significantly with hazard but the correlations are not as high. These include percent members performing (r=.33), percent time spent (r=.35), and task delay tolerance (r=.-34). The negative relationship with delay is due to the fact that the delay scale is inverted with a rating of one rather than nine being the most critical. The lowest and only nonsignificnat correlation was hazard potential with difficulty (r=.-04). Clearly, difficulty and hazard potential do not appear to be linearly related for this specialty.

Data Analysis

A factor printout program (FACPRT) was run to produce all of the tasks sorted in descending order of hazard potential according to the supervisory ratings. The task that the supervisors agreed was the most hazardous was "arm or dearm aircraft armament systems other than guns". An extract from the hazard potential FACPRT listing is given in Appendix D.

The factor printout listing reflects the opinions of the people working in the field and would be highly useful for training designers. However, it must be noted that some of the ratings may have been affected by the rater's knowledge of accidents that had already occurred on certain tasks.

To take the research a step further, prediction models were considered. As mentioned earlier, a major difference exists between this study and other safety research in that this study is predicting tasks that will have accidents occur while the task is being performed rather than predicting who will have an accident while performing the tasks. Consideration was given to predicting the probability of an accident occurring if the task is performed once. In order to predict probabilities it would be necessary to have frequency of performance data for each task; these data are not available. Since a considerable data collection effort would be required to obtain these data, the model predicting probability has been deferred at this time.

From the possible criteria available for analysis, frequency of occurrence of accidents per task was considered the most appropriate. The distribution of the criterion for this specialty was badly skewed. Of the 20 tasks associated with accidents, 11 tasks had only 1 accident occurrence, 3 tasks had 2 accidents, 2 tasks had 3 accidents, 1 task had 4 accidents, 2 tasks had 6 accidents, and 1 task had 10 accidents.

Three models were developed to investigate the relationships among three primary predictors, hazard potential, an index of percent time spent, 1-48 months, percent members performing 1-48 months; six generated variables; and the frequency criterion. The models will be referred to as full, exposure, and hazard. Table 1 illustrates the variables in each model. In addition, the relative contribution of the hazard potential rating was evaluated.

TABLE 1. VARIABLES INCLUDED IN EACH OF THE THREE PREDICTION MODELS

Variables	Full	Exposure	Hazard
Hazard Potential	X		X
Percent Members Performing 1-48 mos.	X	X	
Percent Time Spent 1-48 mos.	X	X	
Hazard Squared	X		X
Members Squared	X	X	
Time Squared	X	X	
Hazard X Time	X		
Hazard X Time Squared	X		
Hazard Squared X Time Squared	X		

Results

The full model predicting frequency of occurrence had an $R=.70$ ($p<.001$); the exposure model with the percent time and percent members variables had an $R=.68$ ($p<.001$); the third model with hazard and hazard squared had an $R=.38$ ($p<.001$). Considering the three primary predictors, hazard potential, the index of percent time spent, and percent members performing; the index of percent time spent on a task accounted for the most variance in the regression models. Percent time correlates .42 with frequency, whereas hazard potential only correlates .27. However, hazard potential does contribute significantly to the full model.

A predicted number of accidents based on the regression weights derived from each of the three models (full, exposure, hazard) with frequency of accidents as the criterion has been computed for each of the 527 tasks for each model. Each of the three sets of predicted numbers of accidents was ordered in factor printouts from the task with the highest predicted number through the task with the lowest predicted number. Appendix F, G, and H are tables showing the cumulative percentage of accidents occurring at different cumulative percentages of tasks.

A chi square was run on each of the sets of predicted number of accidents to test the hypothesis that the distribution of actual accidents over predicted scores was no better than chance. The accident distribution was found to be significantly different from chance ($p<.01$) for each set of predictors. A chi square for independence among the three sets of predicted scores was significant ($p<.05$). In addition, a chi square for independence between the full model and the exposure model was significant ($p<.05$). However, no difference between the full model and hazard was found. Appendix I presents the chi square models. Although the regression model indicates that the hazard potential ratings add significantly to the prediction, the chi square analysis, a somewhat less powerful test, does not indicate significantly different distributions between the full model and the hazard model. A decision to use or not use the hazard potential ratings would be based on further test results together with the expense in money and time involved in collecting the data.

Conclusions and Future Directions

The results of the analysis applied to the aircraft armament speciality are encouraging. One of the most useful products generated is the factor printout of the hazardous tasks as rated by the supervisors in the career ladder. This is an easily understandable tool which could be used by the training designers. The full regression model that was developed to predict expected frequency of accidents accounts for 49% of the variance in this particular speciality. However, it is not yet known how well the model will hold up on cross validation.

Efforts are continuing to determine if the methods so far developed in the present study are valid and generalizable. To that end, research is currently in progress to cross validate and cross apply results developed in the present study.

Additional survey of the 462X0 ladder has been conducted and is under analysis. The survey was performed to collect field recommended training emphasis judgments. Field recommended emphasis ratings are a measure of a task's recommended formal training emphasis, either school or on-the-job, based upon the ratings of field supervisors. The interrelationship of this variable with others already collected will be investigated.

A cross validation study is planned for the 462X0 ladder. When enough new accident data (18 months worth) have been collected, the weights from the frequency of performance model will then be applied to the new data to determine how well the equation would predict in the cross-application.

The method used for analysis of the aircraft armament specialty will be repeated for two additional career fields. Surveys are currently in the field for Fire Protection (571X0) and Fuel (631X0). Results from these two fields may indicate whether the methods developed have any applicability across specialties.

In general, the preliminary findings from this feasibility study have been encouraging. The approach and the methods for predicting tasks which will have accidents are promising. However, results from this initial study must be regarded with reservations until a cross-validation of 462X0 is finished and the results of cross-applications to additional career fields are available.

APPENDIX A: HAZARD POTENTIAL RATING SCALE

Rating Scale	Hazard Potential
1	Extremely Low Hazard Potential
2	Very Low
3	Low
4	Below Average
5	Average
6	Above Average
7	High
8	Very High
9	Extremely High Hazard Potential

APPENDIX B: RATER AGREEMENT INDICES AND AVERAGE MEAN RATINGS FOR TASK FACTORS

Task Factor	R_{kk}*	Average Mean Rating
Hazard Potential	.9315	1.87
Consequences of Inadequate Performance	.9390	6.16
Task Delay Tolerance	.8914	4.52
Task Difficulty	.9302	4.07

* Rater agreement indices for a sample size of 50 raters as estimated by the Spearman Brown formula.

APPENDIX C: CORRELATIONS OF VARIABLES (N=527 TASKS)*

	Hazard Potential	Consequences of Inadequate Performance	Task Delay Tolerance	Task Difficulty	Percent Members Performing	Percent Time Spent
Hazard Potential	1.0000					
Consequences of Inadequate Performance	.6992	1.0000				
Task Delay Tolerance	-.3394	-.5958	1.0000			
Task Difficulty	-.0439	.2711	-.1438	1.0000		
Percent Members Performing 1-48 mos.	.3302	.2870	-.4453	-.2526	1.0000	
Percent Time Spent 1-48 mos.	.3509	.2377	.4431	-.2820	.9702	1.0000

* Correlations above .088 are significant at the .025 level.

APPENDIX D: PRIORITIZED JOB TASKS IN TERMS OF HAZARD POTENTIAL RATINGS

			Seq #	Haz Pot	% Mem 1-48	% Time Spent 1-48	Con of Inad. Per	Del Tol	Dif	Num of Acc
F	162	Arm or Dearm Aircraft Armament Systems Other Than Guns	1	6.2	62	1.9	7.5	1.9	3.8	6
F	170	Load or Unload Non-Nuclear Munitions on Aircraft or Pre-Load Stands or Racks	2	5.9	57	1.6	7.8	2.5	4.2	10
F	172	Load or Unload Preloaded Non-Nuclear Munitions on Aircraft	3	5.9	34	.8	7.8	2.6	4.2	0
F	174	Perform Functional Checks or Tests on Aircraft Armament Circuits While Loading	23	4.3	60	1.7	7.7	2.10	4.21	6
P	426	Drive Ammunition Loaders	76	3.1	18	.4	5.4	4.7	3.1	0
H	230	Perform Operational Checks of Jettison or Emergency Release Systems Using Meters or Indicators	84	2.9	48	.9	7.3	3.6	4.2	1
P	445	Perform Facility Cleaning on Vehicles	396	.8	7	.9	5.9	3.5	4.3	0
			$\overline{X}$ =	1.87	12.78	.19	6.16	4.52	4.07	
			SD =	1.24	11.16	.27	.86	.81	.55	

APPENDIX E: CORRELATIONS BETWEEN PREDICTORS*
AND CRITERION (N=527 TASKS)

	Accident Frequency
Hazard Potential	.2721
% Members Performing (1-48 mos)	.3386
% Time Spent (1-48 mos)	.4215

* All correlations significant ($p<.025$)

APPENDIX F: CLASSIFICATION OF PERCENTAGES OF ACCIDENTS OCCURRING ON DIFFERENT PERCENTAGES OF TASKS ORDERED ON PREDICTED NUMBER OF ACCIDENTS BASED ON FULL REGRESSION MODEL

Percentages of Accidents	Percentages of Tasks
45	1
53	5
67	10
86	20
86	30
90	40
98	50
.	.
.	.
.	.
.	.
100	100

APPENDIX G: CLASSIFICATION OF PERCENTAGES OF ACCIDENTS OCCURRING ON DIFFERENT PERCENTAGES OF TASKS ORDERED ON PREDICTED NUMBER OF ACCIDENTS BASED ON EXPOSURE MODEL

Percentages of Accidents	Percentages of Tasks
45	1
49	5
49	10
67	20
69	30
82	40
92	50
.	.
.	.
.	.
.	.
100	100

APPENDIX H: CLASSIFICATION OF PERCENTAGES OF ACCIDENTS OCCURRING ON DIFFERENT PERCENTAGES OF TASKS ORDERED ON PREDICTED NUMBER OF ACCIDENTS BASED ON HAZARD MODEL

Percentages of Accidents	Percentages of Tasks
33	1
49	5
65	10
88	20
88	30
88	40
96	50
.	.
.	.
.	.
.	.
100	100

APPENDIX I: CHI SQUARES

1. Chi Square for Difference From Chance for Full Model

	Percentage of Tasks				
	20	40	60	80	100
Number of Accidents	42	2	4	0	1

$x^2 = 133.14\ (p<.01)$

2. Chi Square for Difference From Chance for Exposure Model

	Percentage of Tasks				
	20	40	60	80	100
Number of Accidents	33	7	5	2	2

$x^2 = 70.49\ (p<.01)$

3. Chi Square for Difference from Chance for Hazard Model

	Percentage of Tasks				
	20	40	60	80	100
Number of Accidents	43	0	5	1	0

$x^2 = 142.33\ (p<.01)$

4. Chi Square for Independence Among Three Models

		Percentage of Tasks				
		0-1	2-5	6-10	11-20	21-100
Number of Accidents	Full	22	4	7	9	7
	Exposure	22	2	0	9	16
	Hazard	16	8	8	11	6

$x^2 = 19.30$ $(p<.05)$

5. Chi Square for Independence Between Full and Exposure Models

		Percentage of Tasks				
		0-1	2-5	6-10	11-20	21-100
Number of Accidents	Full Model	22	4	7	9	7
	Exposure Model	22	2	0	9	16

$x^2 = 11.19$ $(p<.05)$

6. Chi Square for Independence Between Full and Hazard Models

		Percentage of Tasks				
		0-1	2-5	6-10	11-20	21-100
Number of Accidents	Full Model	22	4	7	9	7
	Hazard Model	16	8	8	11	6

$x^2 = 2.62$ (NS)

REFERENCES

Askren, W. B., Campbell, W. B., Seifert, D. J., Hall, T. J., Johnson, R. C., & Sulzen, R. H. *Feasibility of a computer simulation method for evaluating human effects on nuclear systems safety*. AFHRL-TR-76-18. Wright-Patterson Air Force Base, TX: Advanced Systems Division, Air Force Human Resources Laboratory. May 1976 AD-A025 310.

Christal, R. E. & Weissmuller, J. J. *New CODAP programs for analyzing task factor information*. AFHRL-TR-76-3. Lackland Air Force Base, TX: Occupational and Manpower Research Division and Computational Sciences Division, Air Force Human Resources Laboratory. May 1976 AD-A026 121.

Mead, D. F. *Determining training priorities for job tasks*. Paper presented at the 17th Annual Conference for the Military Testing Association, U. S. Army, Indianapolis, IN, 16-19 September 1975.

Ruck, H. W., Thompson, N. A. & Thomson, D. C. *The collection and prediction of training emphasis ratings for curriculum development*. Paper presented at 20th Annual Conference of the Military Testing Association, U. S. Coast Guard, Oklahoma City, OK, 30 Oct-3 Nov 1978.

Methods for Collecting and Analyzing Task Analysis Data

by

A. John Eschenbrenner
Philip B. DeVries

McDonnell Douglas Astronautics Company
St. Louis, Missouri

and

Hendrick W. Ruck

Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper
are those of the authors and are not necessarily
those of the United States Air Force.

Since the U.S. Air Force (AF) developed its first major instructional system in 1965, the systems approach to training has received considerable emphasis within the Department of Defense and in the civilian sector. The issuance of AF Manual (AFM) 50-2, Instructional System Design, and AF Pamphlet (AFP) 50-58, Handbook for Designers of Instructional Systems, witnessed a realization on the part of the AF that application of modern instructional technologies might yield substantial improvements in the effectiveness and efficiency of AF training programs. In both documents, considerable emphasis has been placed upon achieving close correspondence between training program content and job performance requirements.

The Occupational Survey (OS) is an information source useful for accomplishing job analysis and specifying job performance requirements within the context of AF technical training. However, it does not, nor was it intended to, generate the kinds of data about job performance subtasks or elements and supporting skills and knowledges that are required to design instruction. These data are the products of a rigorous task analysis. The process by which a skilled instructional designer identifies the major procedural steps and makes inferences about skill and knowledge requirements is not well articulated. Additionally, those in the Air Training Command (ATC) who are responsible for conducting and documenting task analyses are Subject-Matter Specialists (SMSs), not educational technologists. The implementation of a simplified task analysis procedure/documentation system and a computer-based task analysis data bank may offer significant economies in the design and revision of technical training courses. A standardized task analysis procedure would help insure that course content decisions are made on the basis of job performance requirements as moderated by training situation constraints; and a computer-based data bank would provide a means of storing, retrieving, updating, and disseminating detailed task analysis information. Ultimately, these economies might be expected to manifest themselves in the form of more effective and less costly training.

The primary objective of this study is to develop and field test a simple-to-use, reliable, valid, and cost-effective/time-efficient task analysis procedure for application by ATC training development personnel responsible for the design and conduct of technical training courses. A secondary objective is to make recommendations regarding the feasibility and utility of implementing a computer-based task analysis data bank and to submit a preliminary data bank design for consideration. End items include:

a. A handbook detailing a standard task analysis procedure that will provide an acceptable degree of uniformity and quality control over task analysis efforts at ATC Technical Training Centers (TTCs); and

b. A systems analysis of present and future AF task analysis requirements, with special emphasis on the recommendations regarding future plans for a task analysis data bank.

The investigative approach employed in this study is straightforward and comprehensive. In Phase I, task analysis procedures currently in use at ATC TTCs were characterized and evaluated, and recommendations for improving the task analysis effort were proposed. In Phase II, a standard procedure was specified and a prototype handbook was developed. It will be field tested at ATC TTCs, and revised on the basis of field test results. In Phase III, the task analysis handbook and the description of data bank requirements will be prepared, reviewed in conference with intended users and management personnel, and revised as necessary prior to finalization. Inherent in this approach is the assumption that continuous involvement of ATC training development and management personnel in the design, testing, and revision process will insure that the final product is useful and will maximize the probability that it will be accepted and implemented.

Phase I

Survey Procedures

A semistructured research interview was employed to gain insight into the task analysis procedures currently being utilized in the AF technical training community. Specific areas of inquiry were the relative percentage of time spent revising existing courses versus developing new ones; procedures currently utilized to accomplish, document, and validate subtask and skill/knowledge analyses; and familiarity with and judged adequacy of the task analysis guidance provided in AFM 50-2 and AFP 50-58.

The sample of interviewees included a full range of training development personnel, including military and civilian education specialists. Instructional System Development (ISD) technicians, and master instructors, who had been or were currently involved with task analysis efforts at the five ATC Training Centers. In addition, training development personnel from the 3306th Test and Evaluation Squadron at Edwards AFB; the School of Aerospace Medicine, Brooks AFB; and the School of Health Care Sciences, Sheppard AFB, were also interviewed.

Results

We found that task analysis procedures and documentation methods utilized at the TTCs were widely variant. Documentation produced in response to inquiries regarding how the results of task analyses were recorded ranged from Plans of Instruction (POIs) to fairly detailed ISD worksheets, most of which were locally designed. It was our feeling that quality control of the task analysis effort across branches within the same group would have been difficult, at best. An integrated quality control program across TTCs would be virtually impossible. Therefore, an attempt to develop and implement a standardized task analysis procedure/documentation system for application at all TTCs seemed a worthwhile pursuit.

We also noted with some interest that no individual or group of individuals at the TTCs was ultimately accountable for the task analysis effort. The issue of accountability is, of course, closely related to that of quality control. For ATC to realize the maximum benefits associated with implementing a standardized task analysis procedure/documentation system and to insure a rigorous quality control program, an articulated accountability system must be defined and implemented.

Frequently heard comments regarding currently available ISD and task analysis guidance documents (AFM 50-2 and AFP 50-58) included, "too complex," "require too much paperwork," and "most applicable to the design of new courses." With regard to the final comment, in response to a direct survey inquiry, we found that training development personnel currently spend the great majority of their time (in excess of 95%) completing task analyses in support of the redesign of existing technical training courses. There seems to be a legitimate need for a simplified task analysis procedure/documentation system that can be applied in the revision of existing courses as well as the development of new courses. Additionally, a preliminary assessment of the feasibility of implementing an automated storage/retrieval system for task analysis data seems warranted. This type of data bank would facilitate the revision of existing courses and would support implementation of a quality control/accountability system.

319

Recommendations/Actions

Based on survey findings and our observations, we recommended that a simplified task analysis procedure/documentation system, including improved procedures for in-process review of task analysis efforts, be developed and field tested at ATC TTCs. Additionally, we recommended investigating the feasibility of providing an automated storage/retrieval system for task analysis data. The Air Force Human Resources Laboratory (AFHRL) and ATC directed us to proceed with the development of a prototype task analysis handbook. Further, ATC agreed to support field testing of the prototype handbook at the TTCs.

Phase II

Handbook Development

The task analysis handbook addresses the design and revision of technical training courses and presumes the existence of a comprehensive task listing in the form of a Specialty Training Standard (STS) or Course Training Standard (CTS). The handbook task analysis procedure represents a best-mix of procedures contained in existing documents and literature, while incorporating the comments and suggestions made by training development personnel during the Phase I survey.

The handbook presents task analysis as a three stage process. Figure 1 presents, in flowchart format, an outline of the handbook task analysis procedure.

Stage A consists of converting STS/CTS task and knowledge items into Preliminary Criterion Objectives (PCO). In Stage B, each PCO is examined and broken down into its component subtasks. Finally, Stage C consists of determining the skills and knowledges which underlie or support each subtask. It was our strong feeling that identification of supporting skills and knowledges had to be addressed if the task analysis effort was to achieve its two primary objectives: (1) insuring that only "need to know" content was included in a course, and (2) providing an adequate information data base to support preparation of objectives and test items. Stage C is primarily inferential in nature and therefore less amenable to proceduralization than other parts of the process. However, some guidelines are offered in support of Stage C activities.

It should be noted that there are a number of differences between the handbook procedures and the detailed task analysis guidance presented in AFP 50-58. First, the task analysis guidance in AFP 50-58 is fragmented, whereas the prototype handbook presents task analysis as an integrated process. Second, AFP 50-58 calls for a considerable amount of task analysis activity prior to finalization of the training standard, while the handbook assumes a training standard as the point of departure. Third, the handbook specifies a single format (the flowchart) for intermediate documentation, and a single form for final

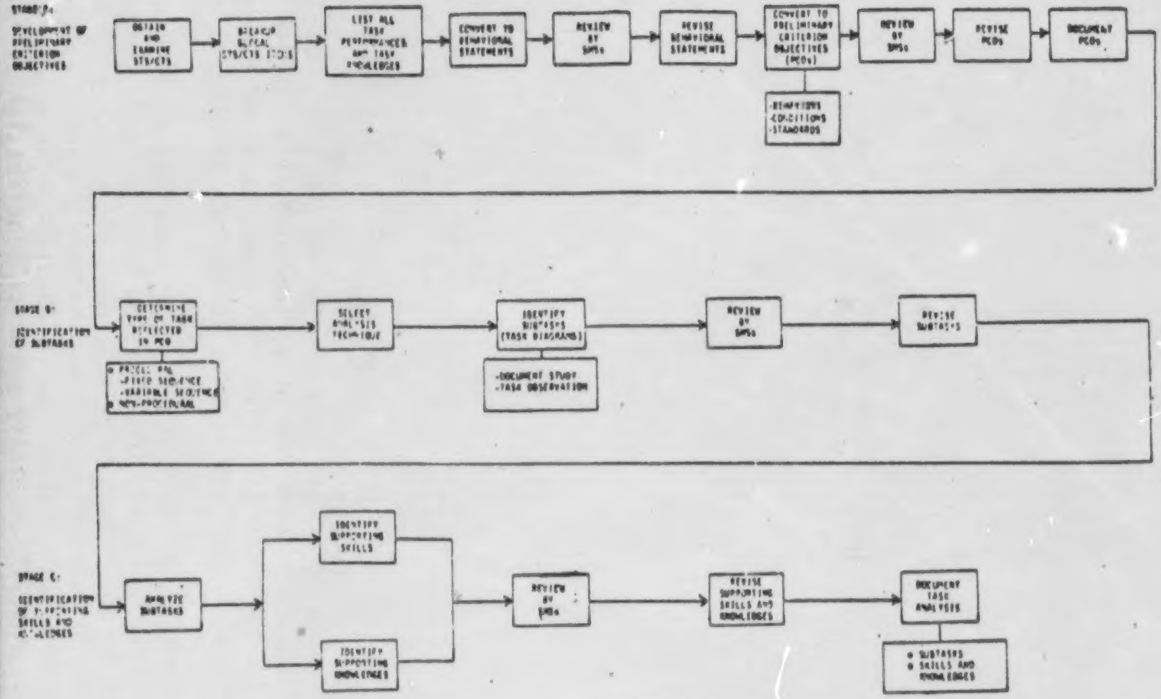

STAGE A:
DEVELOPMENT OF PRELIMINARY CRITERION OBJECTIVES
OBTAIN AND EXAMINE STS/CTS
BREAKUP GLOBAL STS/CTS ITEMS
LIST ALL TASK PERFORMANCES AND TASK KNOWLEDGES
CONVERT TO BEHAVIORAL STATEMENTS
REVIEW BY SMEs
REVISE BEHAVIORAL STATEMENTS
CONVERT TO PRELIMINARY CRITERION OBJECTIVES (PCOs)
-BEHAVIORS
-CONDITIONS
-STANDARDS
REVIEW BY SMEs
REVISE PCOs
DOCUMENT PCOs
STAGE B:
IDENTIFICATION OF SUBTASKS
DETERMINE TYPE OF TASK REFLECTED IN PCO
• PROCEDURAL
-FIXED SEQUENCE
-VARIABLE SEQUENCE
• NON-PROCEDURAL
SELECT ANALYSIS TECHNIQUE
IDENTIFY SUBTASKS (TASK DIAGRAMS)
-DOCUMENT STUDY
-TASK OBSERVATION
REVIEW BY SMEs
REVISE SUBTASKS
STAGE C:
IDENTIFICATION OF SUPPORTING SKILLS AND KNOWLEDGES
ANALYZE SUBTASKS
IDENTIFY SUPPORTING SKILLS
IDENTIFY SUPPORTING KNOWLEDGES
REVIEW BY SMEs
REVISE SUPPORTING SKILLS AND KNOWLEDGES
DOCUMENT TASK ANALYSIS
• SUBTASKS
• SKILLS AND KNOWLEDGES

FIGURE 1 TASK ANALYSIS PROCESS

Figure 1

Outline of Handbook Task Analysis Procedure documentation. Importantly, too, the final documentation form is considerably simpler than the one presented in AFP 50-58. Fourth, the handbook is built on the assumption that task analysis will be performed by a Subject-Matter Specialist (SMS) who is relatively inexperienced in ISD theory and practice. Handbook procedures, therefore, do not require the SMS to conduct an instructional analysis. The procedures in AFP 50-58, on the other hand, call for the analyst to classify each knowledge being analyzed (e.g., chaining, associating) to determine proficiency levels for supporting skills and knowledges, and to specify the amount of practice required to reach proficiency. It is our feeling that these decisions are better left to instructional design specialists. Fifth, and finally, the handbook specifically calls for a series of reviews by SMSs at key points in the task anlysis process. The interaction between analysts and reviewers should provide an excellent safeguard against overtraining. It was our feeling that SMS review and verification is given insufficient emphasis in AFP 50-58.

Table 1
Differences Between AFP 50-58
and Task Analysis Handbook Procedures

AFP 50-58	Task Analysis Handbook
Fragmented Procedures	Integrated Process
Analysis Prior to Finalization of Training Standard	Assumes Existence of Training Standard
Complex Documentation	Simple Documentation
Assumes Instructional Design Expertise	Assumes Technical Subject Matter Expertise
Requires Managerial Review	Requires More Interaction and Review by Other SMSs during Task Analysis

Field Test Procedures

A two-stage field test of the prototype task analysis handbook will be conducted. Stage 1 consists of preliminary tryouts, while Stage 2 will be devoted to feasibility testing (i.e., formal evaluation).

Stage 1 Procedures and Results. Preliminary tryouts were accomplished to obtain information useful for revising the prototype handbook. The goal was to develop an empirical data base that could be used to identify required revisions and make the handbook as useful as possible. A potentially important by-product of preliminary tryouts was a set of task analysis examples directly relevant to AF technical training. The three sites selected for preliminary tryout of the handbook were Keesler AFB, Edwards AFB, and Chanute AFB.

Every attempt was made to insure that the courses chosen for preliminary tryouts of the task analysis procedure encompass the full range of technical training. Basic and advanced training for "soft" skill courses, operator training courses, and maintenance courses were represented. Two or more courses per site were utilized as test beds. For each course, a duty area was selected, and a task performance item and a task knowledge item from within that duty area were chosen for analysis. For each course at each site, two SMSs participated in the preliminary tryout. One of these SMSs served as an analyst, the other served as a reviewer.

Analysts then employed the handbook procedures to analyze one task performance item and one task knowledge item from the selected duty area and documented the analyses. Task analysts were encouraged to ask questions, identify problems, and present suggestions for improving the procedure. If the analyst failed to understand an explanation, another wording or elaboration was attempted. If the analyst failed to understand an example, verbal clarification was provided. Problems encountered, explanations and additional information provided, and suggestions for improvement, as well as typographical errors and other kinds of difficulties that the analysts encountered, were recorded. Reviewers had two tasks during preliminary tryouts. Their primary task, of course, consisted of reviewing task analysis worksheets and documentation. A secondary function involved critically reviewing the handbook in an attempt to identify faulty wording, unclear passages, inadequate explanations, poor examples, improper sequencing, poor layout, typographical errors, and other difficulties. Additionally, general suggestions for improving the handbook and procedures described therein were solicited.

Additionally, each Technical Training Group (TTG) at each TTC provided a senior review team, consisting of an educational specialist and a senior SMS, which examined the handbook, identified problems and made suggestions for improvement, and completed a free-response questionnaire containing items related to the adequacy and practicality of the task analysis procedure/documentation system described in the handbook, as well as items related to appropriateness of style and format.

To reiterate, the objective of the preliminary tryout phase of field testing was to gather information that could be utilized to "fine tune" the handbook prior to feasibility testing (i.e., formal evaluation). At each test site, those training development personnel who participated as task analysis teams and as senior review teams generated a sizable number of suggestions for improving the handbook. There was at each test site substantial overlap between the suggestions put forth by the two groups of participants. In our view, this close agreement constituted consensual validation and provided sufficient justification for revising the handbook in accord with the suggestions made. Not surprisingly, both the number of new and the number of major suggestions generated decreased steadily from site to site. We concluded that the preliminary tryouts had indeed served their primary purpose--a considerable amount of "fine tuning" had been accomplished.

Stage 2 Procedures. The three sites to be used for feasibility testing (formal evaluation) will be Lackland AFB, Lowry AFB, and Sheppard AFB. Three courses per site will be identified as test beds. For each course, a duty area will be selected, and a task performance item and a task knowledge item from that duty area chosen for analysis.

For each course at each test site, four SMSs, one senior SMS, and one training specialist will participate in the feasibility testing. The pool of four SMSs will be divided into two two-person task analysis teams. On each team, one SMS will serve as the analyst, the other as reviewer. The senior SMS and the training specialist will serve as a task analysis evaluation team.

Analysts will utilize the task analysis handbook to analyze the selected task performance item and task knowledge item from the chosen duty area and document the analyses. Those participants designated as reviewers will participate in the analysis and documentation activities in the manner prescribed in the handbook. The amount of time required by each team to complete each major step in the analysis will be recorded. Upon completing the analysis, each analyst and reviewer will be asked to complete a Handbook Evaluation Survey. The survey consists of 43 Likert-type items that solicit opinions regarding the task analysis procedures prescribed in the handbook as well as handbook format and style. Additionally, three free-response items are also included to allow respondents to indicate which handbook features they like best and least and to raise important issues not directly addressed in the survey.

Evaluators will then be asked to review the completed task and knowledge analysis and assess each analysis from the standpoints of accuracy, completeness, and overall adequacy as a basis for the development of objectives, the preparation of tests, and the design of instruction. They will also be asked to judge the degree of correspondence between the analyses produced by the two analysis teams.

Additionally, each TTG at each TTC will provide a senior review team, consisting of an educational specialist and a senior SMS, which will examine the handbook, identify problems, and make suggestions for improvement, and complete the Handbook Evaluation Survey.

The simplicity of the handbook procedure will be assessed by examining analyst, reviewer, and senior review team opinions regarding the readability of the manual, the clarity of the explanations offered, the adequacy of examples included, and the appropriateness of the terminology. These data will be gathered with the Handbook Evaluation Survey.

The validity of the handbook procedures will be assessed by examining the opinions of the task analysis evaluation teams with regard to: the accuracy of each analysis; the completeness of each analysis; and the overall adequacy of each analysis as a basis for developing objectives, preparing tests, and designing instruction. An overall rating of the quality of each analysis will also be solicited.

The reliability of the handbook procedures will be assessed by examining the correspondence between analyses for each course (evaluation team judgments), and the consistency of high correspondence across courses. The consistency with which the new procuedres produce high quality results will provide an additional index of reliability.

Summary and Conclusions

An investigative study was undertaken at the behest of ATC to review task analysis methodologies currently in use, to recommend improvements in current procedures, to develop a simple-to-use, reliable, valid, and cost-effective/time-efficient task analysis procedure. In addition, should a successful procedure be developed, the study would examine the feasibility of providing an automated storage and retrieval system for task analysis data.

Results from Phase I of the study included strong recommendations for development of a simple standardized procedure and documentation system, and establishment of accountability for task analyses. Furthermore, it was recommended that the procedure be oriented toward both course revision and initial course development.

A new task analysis procedure was developed to satisfy the ATC requirements. The procedure differed significantly from current AF recommended task analysis procedures in that it is simpler, designed for SMSs, requires streamlined documentation, requires accountability, and is an integrated process. Preliminary tryouts of the prototype task analysis procedures resulted in a handbook that could be formally evaluated. Conclusions about the success of the handbook must wait until the final formal testing has been conducted and evaluated.

References

Air Force Manual 50-2. Instructional System Design. Washington: Department of the Air Force, December 1970.

Air Force Pamphlet 50-58. Handbook for Designers of Instructional Systems (5 Volumes). Washington: Department of the Air Force, July 1973.

Methodology for Selection and Training of Artillery Forward Observers Job Analysis*

by

John B. Mocharnuk
and
Ruth Ann Marco

Engineering Psychology Department
McDonnell Douglas Astronautics Company
St. Louis, Missouri

INTRODUCTION

The U.S. Army Field Artillery School at Ft. Sill, Oklahoma is charged with the responsibility of training artillery officers in all facets of artillery systems performance. One component of this system is the location of enemy targets and subsequent destruction of these targets through direction of fire by an observer located in a forward position in the combat zone, remote from the artillery pieces. The accuracy and rapidity with which the forward observer (FO) is able to perform these tasks have a direct bearing on the outcome of the battlefield situation, i.e., whether enemy targets are destroyed or disabled. With advances in battlefield weapons technology and enemy mobility, the role of the FO has become even more critical. Recently, concern has been expressed regarding the selection of personnel who are best suited to perform these tasks and the requisite training necessary to increase the efficiency and effectiveness of the combat artillery unit.

In response to this concern, a Weapons System Training Effectiveness Analysis (WSTEA) study was conducted by the Directorate of Evaluation at the Army Field Artillery School. That study focused on the forward observer component of the Field Artillery system. Their findings indicated that considerable improvement in the effectiveness of the system could be achieved by improving the accuracy of both target acquisition and location on the part of the FO.

It is clear from the WSTEA report that FO performance is not at the desired level. The WSTEA evaluation revealed that although accurate fire delivery could be achieved, forward observers required an average of 4.7 artillery rounds in adjustment to achieve the desired accuracy. The Army Training and Evaluation Program (ARTEP) standard is three rounds for adjustment prior to firing for effect. Other results of the WSTEA field evaluation showed self-location accuracy and target location accuracy to be below ARTEP standards.

---

*This is based upon research being conducted for the U.S. Army Research Institute for the Behavioral and Social Sciences under Contract DAHC-19-78-C-0025.

Additional studies (Eschenbrenner & Taylor, 1969; Taylor & Eschenbrenner, 1970; Taylor, Eschenbrenner, & Valverde, 1970; Dominque, 1973; Laveson & De Vries, 1973; U.S. Army Combat Development Command, 1968; U.S. Army Field Artillery School, 1975; and Thomas, 1976) suggest the same conclusion reached by the WSTEA team. FOs are not performing at acceptable levels overall and in some cases, performance is so far below acceptable standards that it would severely impair combat effectiveness. In order to upgrade the performance level of the Field Artillery FO, and thereby improve the combat effectiveness of the field artillery subsystem, increased emphasis must be placed on the selection and training of FOs who can demonstrate competence on combat-referenced operational performance measures. This can be achieved by analyzing the forward observer tasks, developing a profile of the effective forward observer, and specifying the correspondence between this profile and valid performance criteria.

The following paper presents the McDonnell Douglas Astronautics Company - St. Louis (MDAC-St. Louis) approach to the development of a methodology for the selection and training of field artillery FOs.

TECHNICAL APPROACH

The MDAC-St. Louis approach to the selection and training of FOs. incorporates a job analysis of current FO job and skill requirements with a training analysis of the FO component of the Field Artillery Officer Basic Course (FAOBC). In the FO Job Analysis, two techniques, task analysis and profile development, have been combined in order to maximize the amount of information available for the decision process in the training analysis phase. The task analysis element will identify the essential skills and knowledges an FO needs to know in his combat role. The profile development will supplement the task analysis with an examination of the critical characteristics, abilities, aptitudes, personalities, education, and personal histories of the successful FOs. Neither of these techniques, task analysis nor profile development, is particularly innovative in its usual context, especially since task analysis, in the classical usage, does involve some elements of trainee characteristic description. However, the combination of task analysis with the type of profile development procedure that is typically the domain of personnel selection will provide the basic standards for FO selection, as well as the information critical to the determination of FAOBC program effectiveness. Additionally, it will furnish the data necessary to suggest improvements to be incorporated into FO training and to upgrade and standardize that program.

Job Analysis

The primary objective of the FO job analysis is the identification of the critical tasks an FO must perform in order to achieve his mission. These essential job elements will be compared with the existing FO training program to determine if all critical tasks are being taught. TRADOC Pamphlet 350-30, Interservice Procedures for Instructional Systems Development; Phase I: Analyze, outlines four basic procedures to be used in the

358

conduct of a job analysis: 1) development of a tentative task list, 2) authentication of the task list, 3) validation of the task list, and 4) identification of subtasks, conditions, cues and standards. Since the present research is not specifically directed toward the development of detailed behavioral objectives and instructional materials, but to the identification of critical skills, our activity is directed to the task level of specificity rather than to the subtasks, conditions and standards level.

The initial task listing was developed by extracting FO and possible FO tasks from pertinent OBC texts and from direct observation of FO training activities. Special emphasis was placed on Gunnery, Map Reading, and Counterfire texts and on graded and ungraded firing exercises. Once the tentative lists were developed, they were consolidated into a list of candidate FO tasks and a preliminary task categorization scheme was developed.

The list of candidate tasks was reviewed with FAOBC instructors from the Gunnery, Counterfire and Tactics departments at the Field Artillery School. At least ten instructors from each department were interviewed for task selection purposes. Because the refinement of a task listing is an iterative process, the list which reflects the inputs of the FAOBC instructors is not considered a final task listing, but will be subject to further review.

The revised FO task list will be reviewed by additional FO instructors and FOs assigned to organic Field Artillery Units. Structured interviews with fifty FOs are scheduled. The interviewees will be asked to evaluate each task for offensive and defensive scenarios in the following theaters: European theater, Far Eastern theater, Middle East and African theaters. Interview data will be augmented by information collected via questionnaires distributed to FOs who have served in Europe, Korea, Vietnam and CONUS. The questionnaires will also include items pertinent to training and profile development.

Profile Development

Profile development will emerge from analytical and statistical examinations of a critical skills and characteristics list for the effective FO and from an assessment of the makeup of the current FAOBC student population. The list of critical skills and characteristics is being developed primarily from the following three sources and procedures: 1) Examination of the prioritized FO task list, 2) interviews with experienced FOs and FAOBC instructors, and 3) questionnaire responses from experienced FOs.

The examination of the prioritized FO task list presumes that certain tasks demand specific, requisite skills and characteristics. Similarly, to operate specialized FO related equipment necessarily demands certain abilities which must be components of the critical skills list and, observing logical sequence, components of the profile. List elements emerging from this process will be further evaluated when interview and questionnaire data sets are complete.

Interviews will be used not only for further refinement of the critical skills and characteristics listing, but for the generation of new elements for inclusion in the critical skills and characteristics lists. Artillery Officers assigned at Ft. Sill, Oklahoma, and officers assigned at other CONUS installations will be interviewed. The interviews with these officers will serve to provide a more diverse sample.

Characteristics and critical skills identified from the above activities are being subjected to further evaluation using questionnaires. Additional elements of a skills and characteristics listing will be directly solicited using the same questionnaire. Descriptive statistics will be compiled for the questionnaire responses and used for further refining of the profile.

A second major component of the profile development activity relates to the development and refinement of the FO Personal Profile Questionnaire and the provisional validation of the profile developed with that instrument. A developmental version of that questionnaire has been administered to FAOBC 12-78. Item analysis on this version will be completed with comparisons of upper and lower criterion group performance along several criterion dimensions. Those include firing scores for individual graded shoots, a combined firing score, gunnery, counterfire, and tactics grades, and the overall OBC grade. The criterion measures will not be available for a few weeks, but some early frequency data from selected questionnaire items are included in the preliminary results section. The training and intermediate criteria will allow the research team to select those items with the greatest potential for discriminating between high and low ability student FOs. Additionally, certain items provide data useful for training development independent of the criteria. Information gleaned from the analysis of the first developmental form will be used to refine the questionnaire. The revised form will be administered to FAOBC 3-79. Analysis of responses to that questionnaire will serve to further improve the profile development device. The development of the profile will also include an evaluation of characteristics of current OBC students reflected in personal data sheets. Variables identified here will be analyzed in conjunction with factors from the FO Personal Profile Questionnaire and the critical skills and characteristics list. A preliminary model of the effective FO will emerge from this analysis activity.

PRELIMINARY FINDINGS

As an example of how the various steps of the job analysis interact with each other and impact the training analysis, we have developed a series of regression equations and summary statistics for selected samples of FAOBC student course grades and personal profile questionnaire responses.

Data collected from students of FAOBC 6-78 were examined as part of a preliminary hypothesis generating activity. More extensive data sets for three separate samples, all considerably larger, are being collected and will be analyzed to evaluate hypotheses generated in this activity.*

The predictor variable data file for each student included age; source of commission (comprised of four dummy variables, Army ROTC, Navy ROTC, Army OCS, and National Guard with Marine PLC as the reference); marital status; college major (composed of the dummy variables, science and math, business, and education, with liberal arts as the reference); and scores on two tests administered at the beginning of OBC, the Mathematics subtest of the Sequential Tests of Educational Progress (STEP) and the nonverbal subtest of the Lorge-Thorndike Intelligence Test. Criterion measures available for this early analysis included firing accuracy scores from two graded shoots; ten subcourse test scores; and a weighted average of these which will, for convenience and clarity, be referred to as the average grade. Three regression models will be presented and their implications discussed.

The first model was constructed using average grade as the dependent variable and allowing the predictors to enter (or exit) from the model according to a stepwise variable selection procedure. The descriptive linear multiple regression model achieved is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_6 X_6 + \varepsilon$$

Table 1 shows β values; the order of variable selection; the value of the statistic, F, when each predictor variable was entered; and changes in R^2 with the addition of variables. The value of R^2 for the model is .489. Although not great, it is suggestive in light of the modest sample size and the preliminary nature of the analysis.

TABLE 1

SUMMARY OF REGRESSION MODEL 1 - AVERAGE GRADE

Variable Descriptor	Variable	β (In Percentage Points)	Increase in R^2	Total R^2	F To Enter
STEP Score	X_1	.147	.281	.281	17.60
Army OCS	X_2	-3.080	.043	.324	2.80
Navy ROTC	X_3	5.561	.027	.351	1.79
Married	X_4	3.517	.045	.397	3.15
Education Major	X_5	5.184	.046	.443	3.39
Business Major	X_6	3.285	.046	.489	3.61
Constant (β_0)		41.542			

---

*The present set included only 47 students for whom an entire data set was available. The authors are fully aware of the limitations imposed by this small sample size, but conclusions are intended as preliminary and to reflect a "data snooping" activity.

The negative effect (β value) of Army OCS when evaluated against a reference variable of Marine Platoon Leader Course (PLC) graduates indicates a Marine/Army difference. This difference is further amplified by the Navy ROTC effect. Virtually, all students in OBC 6-78 who received their commission through Navy ROTC were Marines.

It is not especially surprising that there is such a marked difference between Marine and Army OBC graduates since the Marines, particularly the Marine PLCs, receive a significantly greater amount of pre-OBC training in map reading, land navigation, and terrain association. These three skill areas have been judged to be critical FO tasks by the FO instructors in the task identification step of the FO task analysis. Additionally, the OBC course of instruction presumes prior training in map reading, land navigation, and terrain association in the allocation of time-to-task instruction. However, interviews with FO instructors reveal this assumption to be false. This is supported by the aforementioned data. If the trend identified by this regression equation is confirmed, a recommendation in the training analysis phase of the present research effort might be to pretest on these three tasks to determine those students requiring remedial work.

College major may have a potentially important effect. As indicated by the regression model 1, the effect of college major accounted for over 9 percent of the variance. Because of the restricted sample, the effect should be treated cautiously. Again, if this trend is confirmed in subsequent samples, more definite interpretation could be developed. Presumably, business and education majors may be more involved with form completion, routine procedure following, etc., than the liberal arts or science major,and it is this practice that may account for the difference.

The second regression model examined the radial missed distance of the location indicated by each OBC student serving as the FO on the mobile shoot firing exercise SW. In a mobile shoot, students function as FOs from a vehicle which is moving between individual firing exercises and is sometimes moving during the actual firing exercise. This means that the student must locate and adjust rounds from multiple locations with less opportunity for carefully determined self location than would be the case with a stationary firing exercise. The descriptive linear multiple regression model with radial missed distance as the dependent measure is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (2)$$

Table 2 shows information important for interpreting this model.

Several important features of this regression model should be noted. First, it accounted for only 21 percent of the variability in the data. Second, the magnitude of the βs is large. Third, only variables indicating source of commission entered the model. If one were to take this model seriously,it would imply that Army OCS graduates and Army officers

TABLE 2

SUMMARY OF REGRESSION MODEL 2 - SW RADIAL MISS DISTANCE

Variable Descriptor	Variable	β (In Meters)	Increase in R^2	Total R^2	F To Enter
Army OCS	X_1	416.3	.180	.180	9.88
Army ROTC	X_2	117.5	.021	.201	1.14
Constant (β_0)		153.0			

who completed ROTC do not achieve the level of accuracy in target location that individuals who obtained their commission from other sources achieve. This is an hypothesis which can be examined in future data sets.

It must be pointed out here that radial Miss Distance should logically be the closest approximation of the operational criterion available in the present training environment. Additionally, the FO of the future field artillery team is more likely to be involved in conducting fire adjustment from a mobile position. Recent developments include the development and testing of a Forward Observer vehicle. As such, identification of predictors of this criterion would be potentially more valuable than identifying predictors of certain other factors.

The third descriptive model looked at the linear multiple regression of the predictors on the combined observed fire grade for all OBC graded shoots and the best two of three hasty target location exercises conducted by the Gunnery Department at the FAS (G0211). The model achieved is:

$$Y = {}_{o} + \beta_1\beta_1 + \beta_2X_2 + \ldots + \beta_6X_6 + \epsilon \qquad (3)$$

Table 3 shows pertinent information regarding this model in the same format as previously reported regression models.

TABLE 3

SUMMARY OF REGRESSION MODEL 3 - OBSERVED FIRE GRADE GO-211

Variable Descriptor	Variable	β (In Percentage Points)	Increase R^2	Total R^2	F To Enter
Business Major	X_1	9.231	.153	.153	8.15
Navy ROTC	X_2	11.144	.116	.270	7.01
Education Major	X_3	9.801	.066	.336	4.28
Married	X_4	4.737	.057	.393	3.96
Army OCS	X_5	-4.581	.066	.459	4.99
Large Score	X_6	-0.119	.036	.494	2.81
Constant (β_0)		91.928			

As with the Average Grade, college major and source of commission have an effect on this grade. (Of course this grade is not independent of the average grade, r = .686.) What this suggests is that the effect of major and source of commission influence observed fire grades and not just total course grade.

A developmental form of the Forward Observer Personal Profile Questionnaire was administered to 192 FAOBC 12-78 students at the beginning of training. Their responses on the questionnaire items will eventually be compared with end-of-course and in-course scores to determine an "FO profile." Without the availability of test scores on FAOBC 12-78 students, very little information can be gleaned from this administration of the questionnaire. However, the responses on two questions are of interest in this discussion of preliminary findings.

The first question asked "What was your first branch choice?" Possible responses included: artillery, infantry, armor, combat engineer, finance, adjutant general and other noncombat branch. FAOBC 12-78 students selected as their first choice: 41% artillery, 6% infantry, 8% armor, 6% combat engineer, 3% finance, 8% adjutant general and 28% other noncombat branch. If the categories are collapsed, these responses indicate that 59% chose some nonartillery branch of the Army as their first choice. Of the 59% that chose a nonartillery branch as their first choice, 57% chose a noncombat branch as first. Noncombat branch was the first choice of 39% of the total sample. These data, if this trend is continued in later samples, suggest a possible motivational factor. The question then arises, should only students who want to be in the field artillery combat arms branch of the Army be admitted? At this time, this is not a viable solution. How then, in the course of instruction, do you change this attitude, not necessarily from wanting to be in the field artillery (albeit desirable) but to an attitude of wanting to do well in FAOBC?

The second question dealt with their judgment of the principle factor involved in most failures to hit the target. Possible responses included: a breakdown in communications, inadequate performance by the FO, inadequate equipment, errors on the part of the gun crew, errors in the FDC, and gun error and weather factors. Fifty-eight percent felt that inadequate performance by the FO accounted for most failures to hit the target. Twenty percent thought it was a result in a breakdown of communications; 11 percent gun error and weather factors; four percent no response; four percent inadequate equipment; two percent errors on the part of the gun crew; and one percent errors in the FDC. These responses were given before the students had received any FO training. If they have this attitude prior to training, how then does it affect their motivation to learn, and, secondly, what can be done within OBC to change this attitude?

The motivational issues raised by these two questions only serve to pinpoint areas requiring further analysis. Only if a relationship between these types of questions and the dependent scores is determined can there be any real, substantive discussion of alternatives.

REFERENCES

Domingue, J. C. The U.S. Army tactical fire direction system (TACFIRE) (AIAA Paper No. 73-418). New York, NY: American Institute of Aeronautics and Astronautics, April 1973.

Laveson, J. I. and DeVries, P. B. Forward Air Controller-Tactical Air Command pilot communciation orientation (Final Technical Report MDC E0888). St. Louis, MO: McDonnell Douglas Corporation, August 1973.

Taylor, C.L. and Eschenbrenner, A.J. Forward air controller visual reconnaissance training manual. St. Louis, MO: McDonnell Douglas Corporation, February 1970.

Taylor, C. L., Eschenbrenner, A. J., and Valverde, H. H. Development and evaluation of a Forward Air Controller (FAC) visual training program (AFAL-TR-70-190). Wright-Patterson Air Force Base, OH: Air Force Avionics Laboratory, September 1970.

Thomas, A. S. Ground observer target acquisition capability: Analysis and interpretation of data from two field experiments (AMS AA-TR-125). Aberdeen Proving Ground, MD: U.S. Army Materiel Systems Analysis Activity, June 1976.

U.S. Army Combat Developments Command Experimentation Command. Ground observer probabilities of acquisition/adjustment (USACDCEC Exp. 31.1, Vol. 1). Fort Ord, CA: U.S. Army Combat Developments Command Experimentation Command, September 1968 (NTIS N. AD-841 633).

U.S. Army Combat Developments Command Experimentation Command. Ground observer probabilities of acquisition/adjustment (USACDCEC Exp. 31.1, Vol. 2). Fort Ord, CA: U.S. Army Combat Developments Command Experimentation Command, September 1968 (NTIS No. AD-841 631).

U.S. Army Field Artillery School. Cannon launched guided projectile cost and operational effectiveness analysis (ACN 18812). Fort Sill, OK: U.S Army Field Artillery School, March 1975.

U.S. Training and Doctrine Command. Interservice procedures for instructional systems development, executive summary and model (TRADOC Pamphlet 350-30). Fort Monroe, VA: U.S. Training and Doctrine Command, Headquarters, August 1975.

U.S. Training and Doctrine Command. Interservice procedures for instructional systems development, Phase I: Analyze (TRADOC Pamphlet 350-30). Fort Monroe, VA: U.S. Training and Doctrine Command, Headquarters, August 1975.

Observer Self-Location Ability and its Relationship to Cognitive Orientation Skills

John R. Milligan, Ph.D. and Raymond O. Waldkoetter, Ed.D.
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 3066, Fort Sill, Oklahoma 73503

Twentieth Military Testing
Association Conference
30 Oct - 3 Nov 1978
Oklahoma City, Oklahoma

OBSERVER SELF-LOCATION ABILITY AND ITS RELATIONSHIP TO COGNITIVE ORIENTATION SKILLS

John R. Milligan, Ph.D. and Raymond O. Waldkoetter, Ed.D.

US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 3066, Fort Sill, Oklahoma 73503

INTRODUCTION

The ability of a human observer to locate himself on the earth's surface in relation to other objects or targets on that same surface has widespread military and civilian application; the importance of which is easily overlooked due to the assumption of the skill's uniform existence among individuals. Self-location or spatial orientation ability is often implicitly assumed to exist at levels common to all individuals in land and sea navigation training even though there is extensive evidence to the contrary (Witkin, 1946; Woodring, 1939). There has been an extensive research effort in the area of spatial orientation related to localized brain damage (Ratcliff, Newcombe 1974; Hecacn, Tzortzis, and Masure 1974), sex differences (Cohen, 1977; Maxwell, Croake and Biddle, 1976; Pellrgrini and Empey 1971), age differences (Howard and Templeton, 1966), and race differences (Osborne and Gregor, 1966), but relatively little research has been specific to self-location or geographical spatial orientation and military map training involving target acquisition for indirect fire weapons. The purpose of the exploratory research reported here is to examine self-location abilities, as they relate to cognitive directional orientation, by developing an instrument capable of identifying those who do poorly or do well on such directional tasks.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

Sincere appreciation is expressed by the authors to Dr. Donald O. Weitzman, US Army Research Institute, whose work in this area generated an interest and provided a framework for the authors. Appreciation is also expressed to MAJ D. Nemetz and SFC E. Johnson, US Army Research Institute, Fort Sill Field Unit, for their assistance in data collection.

REVIEW OF LITERATURE

The importance of self-location abilities was demonstrated by the Army's Human Engineering Laboratories in a field test of the field artillery indirect fire system in the early 1970's (Technical Memorandum 24-70). This field test found that over 50% of the error variance in the indirect fire system was attributed to the forward observer's inability to locate the target or himself in relation to the target within acceptable standards. Army Training and Evaluation Program (ARTEP) standards allow a maximum error of 250 meters in target location. Field tests reveal however, that the average target location error is between 500 to 700 meters. This field test although well designed and executed encountered difficulties in controlling nusiance variables which may have influenced the reliability of forward observer performance as the authors noted in that study. The 50% error variance attributed to the forward observer may and probably does overestimate the error variance. There appears, however, little doubt either empirically or logically, that the accuracy of the forward observer largely determines the accuracy of the indirect fire weapons. The rifle marksman's accuracy is affected by the condition of his rifle and the weather conditions but most importantly is determined by his aim or perceptual judgment. With indirect fire weapons, however, the crew doing the firing neither see the target nor calculate adjustments due to weather, distance, etc. These functional tasks are broken down and performed by other team members who in the case of the forward observer may be separated by many miles from the actual guns being fired. The forward observer generally is the only member of the indirect fire team who can actually observe the target being fired upon; he transmits his observations to the fire direction center (FDC) where this information is processed by calculating weather conditions, gun location, type of munition being fired, etc. These calculations are then sent to the gun crew in the form of elevation and deviations which will be set on the guns and the rounds fired. The forward observer observes the impact of the rounds fired and transmits corrections to the fire direction center who in turn recalculate and send new elevation and deviation information to the gun. The essential difference between the perceptual judgment (aiming) used by the rifle marksmanship and the observing done by a forward observer is in the area of what the researchers call "conceptual associating."

The rifle marksman once he has established the range of his target and adjusted the sights on his weapon is faced primarily with a perceptual alignment task in that he must be concerned with the placement of the adjusted and aligned sights upon the target for accuracy. The forward observer on the other hand is faced with the much more complex task of associating a target he can see on a horizonal plane to a military map drawn in the vertical plane. He must be able to analyze the actual terrain from one perspective and interpolate what that terrain looks like when expressed in symbols and from a different perspective. Thus it is primarily a conceptual task requiring extraction and association of information in a form other than that observed.

Kozlowski and Bryant (1977) studied geographical spatial orientation ability in a series of three experiments in an attempt to further investigate individual differences in orientation skills reported in the research literature. The first experiment divided human subjects (N=45) into categories of either good sense-of-direction or bad sense-of-direction. The subjects were then tested to see if what people say about their sense of direction relates to their actual directional and mapping abilities. The first test consisted of pointing to unseen buildings, a map-drawing task, and a pointing to north and nearby cities task. The results of this experiment indicated that the better the self-report of sense-of-direction the better was the orientation performance. Average pointing error was 19.3° (SD=9.5) and 33.2° (SD=14.6) for good and poor sense of direction subjects respectively, $t(43)=3.41$, $p < .01$.

The second experiment in this research was a refinement of the first with the inclusion of additional independent variables. Subjects were given directions, distance, and time estimation tasks. Results indicated that self-reports of sense-of-direction and self-reports of distance-estimation ability are highly correlated; and the better the sense of direction or distance, the smaller the pointing error. The mean pointing error was 10.79° (SD=5.08) for good sense-of-direction people and 25.71° (SD=19.53) for poor sense-of-direction people. The failure of time or distance-estimation performance to correlate with anything was probably due to lack of variation in the performance data according to the authors.

The third experiment attempted to answer the question "How well would self-reports of directional ability be able to predict spatial performance in a novel environment?" A human size maze was used to answer this question in the form of a section of tunnels underneath a dormitory complex. The subjects were lead through the maze once and then traveled the maze as a group for three trials in which performance measures were observed for time, distance, and direction, along with self-reports of the same performance variables after each trial. The researchers found in this study that people with good and poor senses of direction do not differ in their average pointing error, in the accuracy of their estimation of straight line or route distance to the end of the tunnel, or in their estimation of time spent in the tunnels (F ratios < 1). Analysis of the results of these three experiments led the researchers to conclude that far from having an extreme facility at orientation--one that requires little work; the good sense-of-direction people appeared to be more active and put more effort into the tasks.

The group method of traveling through the maze of tunnels may have hidden some significant differences between the two categories. Those with poor sense-of-direction may have simply went along with the good sense-of-direction people. This possibility was acknowledged in the study by citing the findings of Beck and Wood (1976) which suggested personality differences in people who exhibit exploratory behavior "mixers" and those who stay close to a known place in a novel environment, "fixers" which could account for differences observed.

The interpretation of personality or innate differences in the subjects rather than simple learning/experimental differences between good sense-of-direction people and those of poor sense-of-direction can be supported from the literature. Tryon (1939) conducted a series of experiments on maze "bright" and maze "dull" rats and concluded that sensory abilities or simple learning could not account for the observed differences in the rats. Tryon proposed the hypothesis that good maze learners were better at developing directional sets than poor maze learners. This supports the view that high-level cognitive processes rather than simple learning may account for differences in good and poor sense-of-direction people.

The Field Artillery School (FAS) at Fort Sill as a result of the Human Engineering Laboratories analysis of indirect fire systems, previously cited, attempted a further analysis of forward observer performance (ACN 32750, 1977, WSTEA Phase 1a). The FAS used a comparison of two data groups, one consisted of data gathered from officer basic classes and the other was composed of artillery officers from field units. Evaluation of the institutional data consisted of target location, and observed fire scores correlated with map reading scores, number of shoots, and nonverbal tests. Significant correlations were found among all variables except target location and observed fire scores and target location and number of shoots. These results should be accepted with caution, however, due to the fact that large sample sizes such as this (N=1281) insure that even very small correlations will be statistically significant regardless of the meaningfulness of such correlations.

The field test (N=45) analyzed self-location, target location and shoot scores in relation to map reading scores, previous institutional shoot scores, visual acuity, depth perception, nonverbal tests, and number of practice missions. Correlational analysis revealed that only two pairs of the variables were correlated at a significant level, these were: the nonverbal tests with self-location, and map reading scores with field shoot scores. The fact that so few relationships were found to be significant is surprising but must be considered in light of rather severe methodological problems reflected in the study. Although the FAS study failed to show a significant relationship between target location error and observed fire scores the study concluded that accurate target location ability was the primary shortcoming of the forward observer.

Based upon these results the FAS conducted an additional study to analyze the effect of doubling the amount of map reading instruction given. Comparison between groups of students who had their length of map reading instruction doubled to that of control groups revealed no significant differences between the groups. (WSTEA Phase 1c, undated)

The studies reviewed here are suggestive of differences among individuals in spatial orientation, self-location, and target location abilities. Spatial orientation abilities vary with self estimates of spatial orientation ability and are related to later performance on orientation tasks. Experience and training may be related to orientation performance but as of yet have not been clearly demonstrated in the research. All the studies reviewed here have strongly suggested the presence of personality and/or innate differences which may account for differences in performance.

The purpose of the study reported here was to gather additional empirical data on a limited part of spatial orientation abilities. Particularly, the researchers sought information as to the relationships or differences among individuals on self-location abilities and directional orientation abilities. Significant findings of relationships between these two variables were sought by the researchers as an important starting point or pilot study for larger and more comprehensive research designs.

METHOD

The experimenters used a one-way analysis of variance design in which human observers (N=30) were divided into categories of either high or low self-location abilities (median split) on a previously administered practical exercise in which the observer was required to locate his geographical position in relation to his position on a military map. The experimenters then measured the subjects' ability on three tasks: (1) use of a pointing instrument to point the direction to a series of local landmarks familiar to the subjects, (2) use of a pointing instrument to point to a series of cities within the United States, and (3) the subjects were tested with a visual imagery exercise which required the subjects to mentally follow a complex set of directions and then report the direction they were facing at the conclusion and at various points of the exercise.

SUBJECTS

Subjects were 30 male student officers from an officer basic class at the Field Artillery School at Fort Sill, Oklahoma. All students had completed forward observer, and related subject course areas at the time of testing. Self-location scores (percentage correct) were rank ordered for all 118 students. Each student was assigned a number and

15 students were randomly selected from the top half and 15 from the bottom half (median split) of the class.

APPARATUS AND MATERIALS

Two test instruments were used in this study. The first instrument was a 38 inch diameter circular piece of plywood which could be situated on a flat table. The outer edge of this circle had painted the 6400 mils of a military compass in 10 mil increments. Mils were used in this research since this is the measurement unit used on military compasses and can be easily converted to degrees. The center of this circle had a rotating post with a 38 inch pointer which could be pointed in any direction and the direction read in mils off the circular base. Subjects were individually tested in a lighted but enclosed room by showing them the correct direction to true north with the mils and the pointer correctly oriented. Each subject was then asked to move the pointer as close as possible to the actual direction of six local areas in which the student had frequent contact i.e., student mail room, post exchange, etc. Appendix A contains a scoring guide of all locations and their correct directions. The subjects were also required to point the direction to six cities using the pointing instrument thereby providing measures of both local and national geographical orientation.

The second test instrument used in this study was a mental imagery exercise consisting of a single sheet shown to the subjects with square grids covering approximately two-thirds of the page. Individual subjects were asked to close their eyes and imagine themselves at the top of the series of squares or grids facing a specified direction. They were then asked to imagine themselves walking along the grid lines in whatever direction and for whatever distance the experimenter instructed, then at various points along this path they were asked what direction they were facing. Each subject completed three of these mental imagery exercises. Instructions with the plotted paths for each of the three exercises are presented in Appendix B to this paper.

PROCEDURE

Subjects were randomly selected for each of the two groups as previously described and ran individually. The experimenter briefly described the study to each subject and obtained informed consent. Then each subject was taken into a lighted room where the pointing instrument was located. There was no attempt to eliminate directional visual cues within the room. The subject was shown the operation of the pointing instrument and then the instrument pointer was placed on true north and the subject asked to point to the previously described locations.

RESULTS

LOCAL POINTS

One-way analysis of variance was used to evaluate the group differences in pointing to six local areas with which the subjects had daily to weekly contact. Absolute error scores measured in mils from the actual azimuth measured from true north were used in this analysis as the dependent variable. Group assignment was the independent variable with group one consisting of subjects who had scored above the median on a field self-location test and group two consisting of those who had scored below the median on the same self-location test. Table 1 presents the results of this analysis.

Group one (high self-location scores) performed significantly ($p < .04$) better than group two (low self-location scores) on pointing to local points as was expected. Table 2 presents the means, standard deviations and errors for these two groups.

Insert Table 1 and 2 about here

As can be seen from these tables the relative difference is rather small when the mils are converted to degrees (approximately 15° error for group one and 18° error for group two). Although this is a relatively small difference this data provides evidence as to the utility of a pointing instrument in differentiating between high and low scorers in self-location tasks.

DISTANT CITIES

One-way analysis of variance as previously described in the analysis of local points was used to analyze the differences in groups for pointing to distant cities. The results of this analysis are presented in Tables 3 and 4.

Insert Table 3 and 4 about here

As in the previous analysis, significant differences were obtained between groups ($p < .03$) on pointing to distant cities. Again examination of the results of the analysis of variance and means, SD, and SE reveal the pointing instrument was effective in differentiating between groups.

VISUAL IMAGERY

The third analysis as in the first and second revealed significant differences ($p < .002$) between the two groups on the visual imagery tasks.

Insert Table 5 and 6 about here

As can be seen from an examination of Tables 5 and 6 the visual imagery task produced what appears to be the greatest magnitude of differences.

CONCLUSION

The purpose of this study was to examine relationships among self-location abilities and performance on an orientation task requiring estimates of compass directions and geographical spatial orientation using visual imagery. The results of the preliminary research have clearly demonstrated that differences between high scores and low scores on a self-location test can be differentiated by use of a simple pointing instrument and visual imagery task. The results although promising must be accepted with caution due to the relatively small sample size, lack of biographical data on subjects, lack of test retest reliabilities using the instruments, contamination of the criterion variable, relative little variation in the criterion variable, and other uncontrolled variables which may impact upon spatial orientation and self-location skills which were not included in this pilot research. These same cautions, however, provide the foundation for an expanded investigation in which a multivariate statistical design will allow for greater control of variables and analysis of their contributions to performance in self-location and target location abilities.

TABLE 1[a]

Analysis of Variance of Mean Errors in Pointing to Local Points for Groups 1 and 2[b]

Source	SS	df	MS	F
Between Groups Treatment	34884	1	34884	4.60[c]
Within Groups Error	212431	28	7587	
Total	247315	29		

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Unit of measure is in Mils with 6400 mils = 360°.

[b] Group 1 = Subjects scoring above median on self-location test.
Group 2 = Subjects scoring below median on self-location test.

[c] $p < .04$

TABLE 2[a]

Means and Standard Deviations and Errors for Groups[b] on Pointing to Local Points

Group[b]	Mean	SD	Standard Error
1	264	67	17.32
2	332	103	26.68
Total	298	92	16.86

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Unit of measure is in Mils with 6400 mils = 360°.

[b] Group 1 = Students scoring above median on self-location test.
Group 2 = Students scoring below median on self-location test.

TABLE 3[a]

Analysis of Variance of Mean Error in Pointing to Distant Cities for Groups 1 and 2[b]

Source	SS	df	MS	F
Between Groups Treatment	150946	1	150946	5.43[c]
Within Groups Error	778886	28	27817	
Total	929832	29		

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Unit of measure is in Mils with 6400 mils = 360°.

[b] Group 1 = Students scoring above median on self-location test.
Group 2 = Students scoring below median on self-location test.

[c] $p < .04$

TABLE 4[a]

Means and Standard Deviations and Errors for Groups[b] on Pointing to Distant Cities

Group	Mean	SD	Standard Error
1	366	83	21.49
2	507	221	56.98
Total	437	179	32.69

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Unit of measure is in Mils with 6400 mils = 360°.

[b] Group 1 = Students scoring above median on self-location test.
Group 2 = Students scoring below median on self-location test.

TABLE 5[a]

Analysis of Variance of Scores[a] Obtained on Visual Imagery Test for Groups 1 and 2[b]

Source	SS	df	MS	F
Between Groups Treatment	2484	1	2484	11.73[c]
Within Groups Error	5933	28	212	
Total	8417	29		

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Scores represent percent correct

[b] Group 1 = Subjects scoring above median on self-location test.
Group 2 = Subjects scoring below median on self-location test.

[c] $p < .002$

TABLE 6[a]

Means and Standard Deviations and Errors for Groups on Visual Imagery Test

Group[b]	Mean	SD	Standard Error
1	90	12	3.05
2	72	17	4.36
Total	81	17	3.11

Note. N=30; 15 per group. Numbers rounded to nearest whole number.

[a] Unit of measure is in Mils with 6400 mils = 360°.

[b] Group 1 = Students scoring above median on self-location test.
Group 2 = Students scoring below median on self-location test.

APPENDIX A

POINTING INSTRUMENT SCORING GUIDE

Location Name	Location Azimuth
1. Officers Club	5855
2. Main PX	5075
3. Ft Sill Blvd Exit	3610
4. Key Gate	2490
5. Mail Room	1825
6. CF Department	4900
7. Oklahoma City	0710
8. New Orleans	2150
9. Dallas	2550
10. Houston	2670
11. Kansas City, MO	0620
12. Denver	5610

APPENDIX B

VISUAL IMAGERY EXERCISE

NARRATIVE INSTRUCTIONS GRID #1

1. Graphic Representation: See Attached Sheet

2. Scoring Procedure: Score one point for each correct direction given by the subject. Ask the subject for his direction at each place indicated in the narrative.

3. Narration:

    a. Close your eyes and imagine yourself facing South on the grid previously shown to you.

    b. Proceed two blocks South, Stop.

    c. Turn 90° left, now proceed two blocks and Stop.

    What direction are you now facing? (Correct answer is East)

    d. Now turn left 90° and proceed two blocks, Stop.

    e. Turn left 90° and proceed two blocks and Stop.

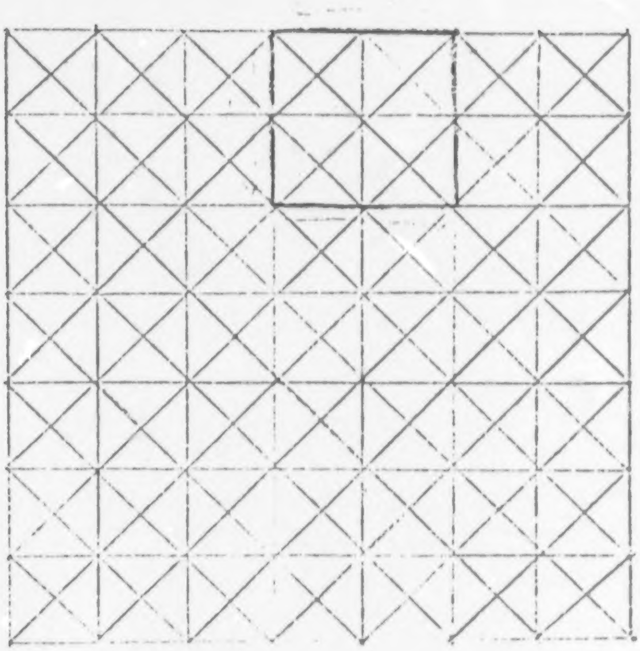
    What direction are you now facing? (Correct answer is West)

    If the subject correctly answers both questions score 2 for this example.

4. Now give the subject a blank grid and ask him to draw the directions he followed in this example.

5. Ask the subject for any questions to clarify the procedure.

6. Proceed to the next exercise if the subject understands the directions.



NARRATIVE INSTRUCTIONS GRID #2

1. Close your eyes. Imagine yourself facing South on the grid you were just shown.

2. Proceed one block and Stop.

3. Turn 90° left, walk one block and Stop.

4. Turn 90° right, walk one block and Stop.

   What direction are you now facing? (A3, South)

5. Turn right 90° proceed one block and Stop.

6. Turn right again 90° proceed one block and Stop.

   What direction are you now facing? (A5, North)

7. Turn right 90° proceed one block and Stop.

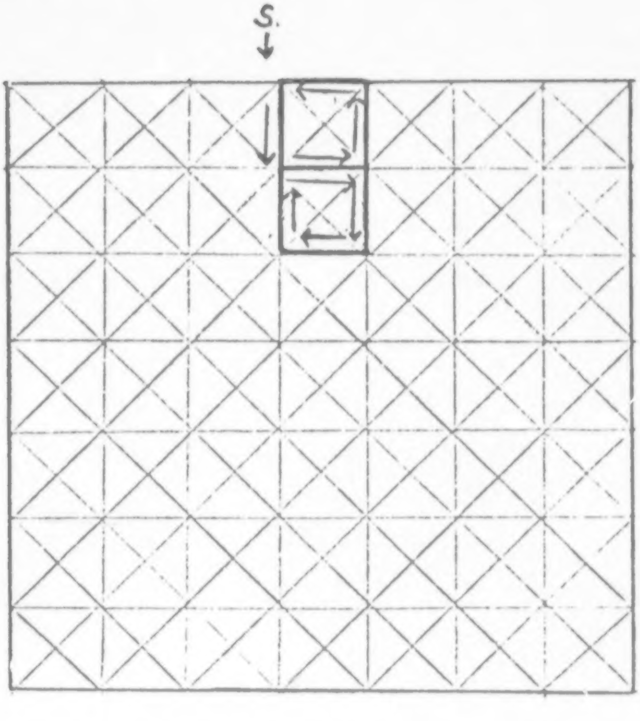
   What direction are you facing? (A6, East)

8. [illegible] left 90° proceed one block and Stop.

9. Turn left 90° proceed one block and Stop.

   What direction are you now facing? (A8, West)

10. On this blank grid page draw the route you have been following.


S.

NARRATIVE INSTRUCTIONS GRID #3

1. Close your eyes. Imagine yourself facing East on the grid you were just shown.

2. Proceed two blocks and Stop.

3. Turn right 90°, now turn 45° more to the right and proceed two blocks and Stop.

   What direction are you now facing? (A2, SW)

4. Turn left 90°, now turn 45° more to the left and proceed two blocks and Stop.

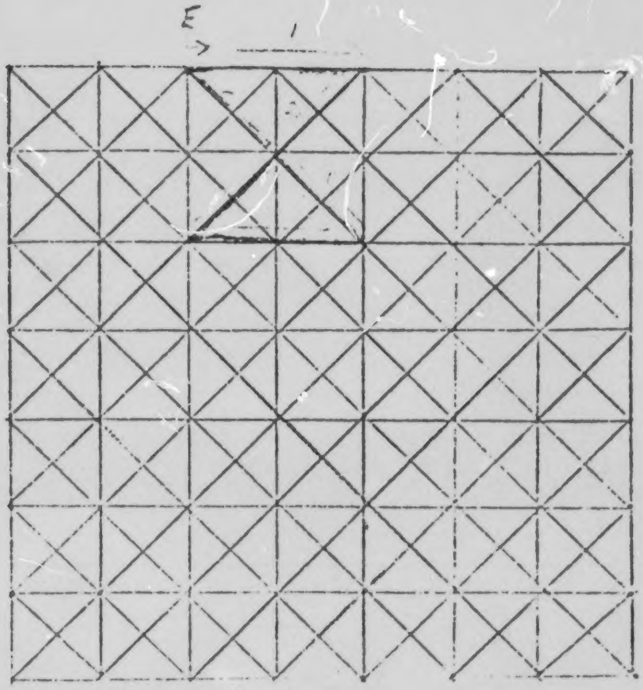
   What direction are you now facing? (A3, E)

5. Turn left 180° then turn right 45°.

   What direction are you now facing? (A4, NW)

6. Proceed two blocks in this direction and Stop. Turn left 45°.

   What direction are you now facing? (W)


E
1

REFERENCES

Beck, R. J. and Wood, D. Cognitive transformation of information from urban geographic fields to mental maps. *Environment and Behavior*, 1976, 8, 199-238.

Cohen, D. Usefulness of the group - comparison method to demonstrate sex differences in spatial orientation and spatial visualization in older men and women. *Perceptual and Motor Skills*, 1976, 43(2), 388-390.

Directorate of Evaluation, US Army Field Artillery School, Fort Sill, OK. *Weapon system training effectiveness analysis - The forward observer phase 1a base line*. ACN 32750, May, 1977.

Directorate of Evaluation, US Army Field Artillery School, Fort Sill, OK. *Weapon system training effectiveness analysis - forward observer*. *Phase 1c*. Undated.

Hecaen, H., Tzortzis, C., Masure, M. Spatial orientation difficulties in a route-finding test by patients with unilateral cortical lesions. *Perception*, 1972, 1(3), 325-330.

Howard, I. P., and Templeton, W. B. *Human spatial orientation*. New York: Wiley, 1966.

Human Engineering Laboratories, Aberdeen Research and Development Center, Aberdeen Proving Ground, Maryland. *Human engineering laboratories battalion artillery test (HELBAT 1)*, Technical Memorandum 24-70, September 1970.

Kozlonski, L. T., and Bryant, K. J. Sense of direction, spatial orientation, and cognitive maps. *Journal of Experimental Psychology*. *Human Perception and Performance*, 1977, 3(4), 590-598.

Maxwell, J., Croake, J., and Biddle, A. Sex differences in the comprehension of spatial orientation. *Journal of Psychology*, 1975, 91(1), 127-131.

Osborne, R. T. and Gregor, A. J. The heritability of visualization, perceptual speed and spatial orientation. *Perceptual and Motor Skills*, 1966, 23(2), 379-390.

Pellegrini, Robert J., Empey, John. Interpersonal spatial orientation in dyads. *Journal of Psychology*, 1970, 76(1), 67-70.

Ratcliff, G. and Newcombe, F. Spatial orientation in man: Effects of left, right and bilateral posterior cerebal lesions. Journal of Neurology, Neurosurgery and Psychiatry, 1973, 36(3), 448-454.

Tryon, R. C. Studies in individual differences in maze learning. VI. Disproof of sensory components: Experimental effects of stimulus variation. Journal of Comparative Psychology, 1939, 28, 361-415.

Witkin, H. A. Studies in geographic orientation. Yearbook of the American Philosophical Society, 1946, 152-155.

Woodring, P. D. A technique for the investigation of direction orientation in human beings. Papers of the Michigan Academy of Sciences, 1939, 24(pt. 4), 147-152.

JOB ANALYSIS IN THE US ARMY
MEDICAL TRAINING ENVIRONMENT

APPROVED FOR
PUBLICATION/PRESENTATION
ACADEMY OF HEALTH SCIENCES
UNITED STATES ARMY
FORT SAM HOUSTON, TEXAS

J. S. Tartell
Academy of Health Sciences
Ft. Sam Houston, TX 78234

The views of the author are his own and do not purport to reflect the position of the Department of the Army or the Department of Defense.

ESTABLISHING THE PROGRAM

Prior to detailing the application of job analysis techniques to the design of enlisted medical training, it would appear appropriate to outline how and why the Army Medical Department came to use the Instructional Systems Development technology which includes the use of job analysis.

As has been the case with its sister services, the U.S. Army has, for many years, been under the scrutiny of the Congress and the Federal Executive Branch. The focus of this scrutiny has been an effort to restructure the training establishment with the intent of changing the student to staff ratio and to make more personnel available for assignment to combat units. There were a number of ancillary issues raised , two of which were the methods of instruction and the cost of the training.

The impact of the Congressional concern was expressed in a legislative amendment to the FY 76 Defense Authorization Bill (House Report No. 94-413) which mandated a study of DOD training establishments. The effect of this legislation on the U.S. Army was to force the issues of modernizing training procedures, streamlining training structures, and minimizing training fund expenditures. During the same year, the DOD in its Report of Training to the Congress, endorsed a new training development model, the Instructional Systems Development (ISD) approach, which subsequently was adopted.

The ISD philosophy was to implement training based upon tasks the trainee would subsequently perform on the job. This philosophy placed the Army Medical Department and the Academy of Health Sciences in a dilemma. The Academy of Health Sciences was committed to enlisted technical training based on the traditional model of education. (The Academy of Health Sciences is the Army Medical Department's only formal school with a staff and faculty of approximately 2150, a resident student population of more than 33,000 annually, and over 30,000 nonresident students enrolled in extension courses.)

To further complicate this dilemma, two other problems arose. The first of these problems lay in the area of gathering sufficient expertise to implement the new training philosophy, task based training, while continuing the on-going training mission to support the needs of the Army. The second problem was the resistance of a largely successful organization to a basic change in both philosophy and organization.

The final catalyst for this monumental change in philosophy and method of operation was the assignment of a new Superintendent to the Academy of Health Sciences. The arrival of a new commander, with a new pragmatic approach, resulted in considerable acceleration of the change process and provided guidance in terms of product oriented direction with a rigid timetable.

With the philosophical decisions made, the next problem was to implement the ISD approach. In order to accomplish this, a task force was established which drew upon the talents and resources available within the Academy of Health Sciences. There were two basic problems in constructing the Task Force. The first was the necessity to continue the traditional training program, thus somewhat restricting the personnel who could be removed from their teaching or administrative positions. The second problem lay in taking the ISD directional pamphlets which were largely a philosophical approach to developing task based training and converting the philosophy into a pragmatic product oriented mode of operation.

These two problems were solved by detailing a number of highly educated personnel to the Task Force and allowing the group approximately thirty days to thoroughly review, digest, and educate themselves as to the vagaries of implementing the ISD philosophy. The self-education process included review of Congressional hearings and documents, technical materials from the ISD model and the history of the ISD process. By November 1976, each member of the Task Force had an overview of how his efforts would fit into the total ISD picture. At that point the Task Force embarked on its first ISD effort, to establish a plan and test the plan by developing a single course of instruction. The specialty chosen for the initial trial effort was the Medical Specialist, MOS 91B, with a target date for course validation of October 1977.

Now, to move from the history of the establishing of the ISD method of training development to the initiation of the job analysis efforts, in April 1977, the Academy of Health Sciences obtained a qualified job analyst. Despite the fact that this occured considerably after the initiation of the Medical Specialist ISD effort, the job analysis procedure was begun. The results were to serve two purposes; first, to establish lines of communications with the Army Occupational Survey Program data base at the U.S. Army Military Personnel Center (MILPERCEN) in Alexandria, Virginia and second, to validate the efforts of the Task Force in establishing a task based training package.

Personnel in the Military Occupational Data Division of MILPERCEN were extremely cooperative in allowing the establishment of informal lines of communications and providing data as rapidly as their system and the U.S. Postal Service would allow. The Academy of Health Sciences was fortunate in having coordinated with the Military Occupational Data Division during the period of 1976 and early 1977 in the construction of job inventory questionnaires for most of the medical specialties. The data to support the ISD efforts had been gathered from September 1976 through April 1977 and much of the data was available for processing. The job analysis of the Medical Specialist, MOS 91B, began in May of 1977 and a final occupational survey report was published in September of 1977. Interaction between the job analyst and the other members of the Task Force led to consideration of survey findings in the development of the new Medical Specialist course.

By October 1977, the Task Force had established a plan for implementing ISD methodology for course development at the Academy of Health Sciences. The plan had been tested in the development of the course of instruction for the Medical Specialist MOS and a recommendation was made and approved to formalize the organization and methodology and continue with the remaining thirty specialties and associated courses of instruction.

JOB ANALYSIS FINDINGS

Since the initiation of the Instructional Systems Development process at the Academy of Health Sciences, occupational surveys have been completed for nine specialties. These specialties are:

Medical Specialist, MOS 91B
Medical Supplyman, MOS 76J
Hospital Food Service Specialist, MOS 94F
Veterinary Specialist, MOS 91R
Behavioral Science Specialist, MOS 91G
Patient Administration Specialist, MOS 71G
X-Ray Specialist, MOS 91P
Clinical Specialist, MOS 91C
Operating Room Specialist, MOS 91D

In addition, analysis of the occupational data for two other specialties are in progress. These are the:

Medical Laboratory Specialist, MOS 92B
Dental Removable Prosthetic Specialist, MOS 42D

In an attempt to illustrate the utility of the occupational survey data, selected findings will be briefly discussed. One of the most valuable contributions that an occupational survey can provide to the training development process is the identification of the different jobs which exist within each specialty and the tasks individuals perform when accomplishing those jobs. For example, the job structure analysis for the Hospital Food Service Specialist, MOS 94F, occupational survey indicated the existance of eleven different jobs within the specialty. The eleven different jobs could be grouped together to form two large clusters of jobs and two smaller separate job classifications. Personnel in one of the large job clusters, titled Food Preparation Specialists, who represented 53 percent of the sample, performed tasks virtually identical to those performed by another specialty, the Food Service Specialist, MOS 94B. On the basis of this information, coupled with additional data, consideration is being given to consolidating the food preparation phase of training for the two specialties at a single location with an additional period of training provided for Hospital Food Service Specialists in the areas of their specialty peculiar to the hospital environment.

A second example of the utility of the job structure information occurred in the Patient Administration Specialist, MOS 71G, occupational survey. The job structure analysis identified eleven separate jobs within the specialty. A number of these jobs were found to be performed by personnel in their second or subsequent enlistments. In reaction to this information, the task analysis team is recommending that training in these areas be given at some time other than in the initial resident course. If such a recommendation is approved, there could be a significant savings in training funds.

However, all occupational survey information must be considered only as a point of departure. Prior to the implementation of any recommendations from an occupational survey, information related to many other factors must be considered. Some of these factors are overall contributions to unit mission, the ability of the individual to perform collective tasks, and the impact on the individual's ability to expand his base of knowledge.

A second aspect of occupational survey information which impacts on training decisions relates to the probability that an individual will perform a task. As an example, in the Medical Supplyman (MOS 76J) occupational survey, there were very few tasks performed by large percentages of survey respondents. The inventory questionnaire included a total of 392 task statements which was a reasonably comprehensive list. The average number of tasks performed by any one respondent was 55, with the average dropping to 46 tasks when the data base was restricted to those in their first enlistment (the target for the initial resident training course). This information, when considered in concert with two other facts; (1) there were three tasks performed by a least half of the target population; and (2) there were an additional sixteen tasks performed by at least one-third of the target population; led to the conclusion that a task based cost-effective training course would be difficult to develop.

A second example of the impact of task performance data on training development came from the occupational survey of the Medical Specialist, MOS 91B. The survey data yielded a rather broad base of tasks which would be appropriate for inclusion in an initial resident training course. There were, however, two substantial problems with this information: (1) Many of the tasks which were performed by personnel at that time were not the ones which would be required to be performed should the individual be placed in a hostile environment (because the Medical Specialist, MOS 91B, is the individual commonly referred to as the Combat Medic; and (2) many of the tasks performed by individuals are not appropriate to include in a specialty training course (these are primarily those tasks related to vehicle maintenance. a responsibility inherent in the job of a soldier).

Another illustration of the impact of occupational survey information is the discovery of the unpopular. These are findings which may be illustrated by the following examples. In the Clinical Specialist (MOS 91C) occupational survey, the job structure analysis identified a small job group (representing approximately four percent of the population) where the personnel were performing tasks which were the same as those performed by a relatively large job group in the Medical Specialist, MOS 91B, occupational survey. This was an unpopular discovery because the Clinical Specialist, MOS 91C, receives approximately one year of training while the Medical Specialist, MOS 91B, receives approximately twelve weeks of training. Another example of an unpopular finding occurred in the Medical Supplyman, MOS 76J, occupational survey. The survey data revealed a differential utilization pattern between the male and female survey respondents. The male respondents performed shipment and storage tasks to a much greater degree than the female respondents, who performed administrative supply tasks to a substantially greater degree.

A final illustration of the impact of occupational survey data lies in the discovery that a specialty can be appropriately described and that training prepares the individual to perform his/her job. The discovery that all is reasonably well within a specialty is too often dismissed while a discovery that something is wrong or in error is trumpeted out of proportion. This impact of the occupational survey information is as important as any other impact and perhaps the most overlooked. In addition, conducting an occupational survey leading to the conclusion that all is well is often not very exciting. The findings, for example, that the Veterinary Specialist (MOS 91R) has a broad and complex job, which included conducting the food inspections for all Army installations under a myriad of regulations and guidelines, was not new to anyone. The finding that an X-ray Specialist, MOS 91P, must be trained to perform a wide range of different radiographic tests was a well-known fact prior to the completion of the occupational survey. However, what is important is that after the completion of the occupational survey, the feelings, intuitions, and preconceived notions can be validated and the training programs can be based on empirically substantiated information.

NOW AND THE FUTURE

The implementation of this new method of training development is well underway. To the present, job analysis has been accomplished for eleven specialties. Task analysis has been completed for five of these specialties and is in progress for an additional three. A new course has been designed and tested for one specialty. The program is clearly still in its infancy. With the development of this new approach to course construction have come many problems, two of which will be discussed.

One of the major areas of concern with the new approach is the relationship between the "what is," as represented by the occupational survey information, and the "what may be," when personnel must perform in a hostile environment. Directly related to this concern is the fact of dealing with the distinctly unique requirements of the medical community. The concept of the "critical task" takes on a very real meaning in a medical emergency. Training programs must be designed to prepare the individual to perform tasks for which the probability of performance may be limited. This requires exposure to the task, not only in the training environment, but also in some form of continuing training beyond the resident course. The use of unit training and Training Extension Courses (TEC) are a partial answer to this problem.

A second area of concern with the new approach involves the cognitive nature of many of the tasks performed by medical personnel. This aspect of task definition and performance became increasingly evident in the development and analysis of tasks for the Behavioral Science Specialist, MOS 91G. Personnel in this specialty deal with individuals who have problems coping with their environment and manifest any number of external and internal abnormal behaviors. The normal task analysis processes (standards, conditions, cues, etc.) were not derived and they are not generally effective in dealing with tasks related to human cognitive skills. In this area the Academy of Health Sciences is developing a supplement to the ISD model to aid in the development of training in the area of cognitive skills.

But what does the future hold for continued implementation of the job analysis effort within the Army medical training environment. The immediate future appears to be relatively well planned with ISD efforts proposed for all of the enlisted medical specialties. These efforts alone will consume the better part of the next three to four years. In addition, there are a number of special projects which illustrate the growth of the ISD program in the medical training community. Such special efforts are; the development of a pre-command course for medical command selectees (what do medical commanders do and what do they need to know?), an attempt to design a front end analysis effort to facilitate the design of a course of training for the Special Forces Aidman (a distinctly different type of medic), the beginning of ISD efforts in the officer arena (a new undertaking in the medical profession), and an assessment of the supervisory and management skills required of commissioned and noncommissioned officers.

CODAP: A NEW MODULAR APPROACH TO OCCUPATIONAL ANALYSIS

By

Michael C. Thew and Johnny J. Weissmuller

INTRODUCTION

The increasing complexity of career fields requires a corresponding increase in the number of task items within an occupational survey. Survey booklets containing 800 to 1200 items are not unusual. Initially, the incumbent was required to read every task item in order to locate those which were relevant to the job. (Appendix A) Because this was an onerous chore, tasks were overlooked and the reliability of the responses could be questioned. By ordering the tasks on some type of commonality, an organization takes place which simplifies the identification of tasks by the incumbent. (Appendix B) This method of organizing the tasks by duties within the job inventory is widely used and works well for data collection. However, not all users find this organization useful when analyzing the data for their particular needs. Recently, methods have been developed to facilitate the reorganization of tasks into new categories called modules. Module definition always occurs after the data base has been generated from the survey instruments.

DATA COLLECTION
- Tasks
- Tasks Within Duties

DATA PRESENTATION
- Tasks
- Duties
- Tasks Within Duties
- Modules

MODULE DEFINITION

The two steps involved in creating modules are definition and assignment. Definition consists of defining the attributes or rules for the organization of tasks into modules. Assignment consists of the application of those rules to the collection of tasks into modules. This is usually done by a person who is judged qualified to decide whether or not a task meets the requirements set by the definition; i.e., a subject matter specialist. When these requirements are quantifiable (measurable by a range of numeric values) an automated approach of combining the tasks may be utilized. Data displayed by user defined modules may provide insights about the survey which are not readily apparent from the original order. This is an extension, not a replacement, for the task, duty or task within duty display formats. The best method is always determined by how the data are going to be used which is an especially important consideration in the module definition.

EXAMPLES

The following examples illustrate a few applications of user defined modules in occupational analysis.

Example 1: Relating Training Requirements to Tasks Performed. Suppose a technical trainer says, "Right now every student is taught how to overhaul engines. We hypothesize that only second term enlistees are doing this job while first termers merely assist with parts of the process. If this is true, we could emphasize the training on those tasks which the first termers actually perform." What the trainer is asking for is a report showing the percent of first termers and the percent of second termers who overhaul engines. Looking at the task inventory list, we find there are no tasks titled Overhauling Engines. Closer examination of the Task Inventory list reveals that several tasks might be associated with engine overhaul. At this point, a subject matter specialist familiar with the operation of overhauling engines is asked to identify which of the tasks in the inventory are applicable. A mark will be placed by those tasks which belong to the new module. (Appendix C) The tasks can now be reorganized into a new pseudoduty or module labeled "Overhauling Engines". The reorganized report of percent members performing data now provides the trainer with information necessary to make his decision. (Appendix D) It is important to note that instead of constructing and administering a new survey, we have decided only to reorganize the existing inventory in a manner that is acceptable to the user's needs. This approach reduces both time and cost.

Example 2: New Task Categories Vs Time-in-service. In another case, someone might ask: "Suppose we separate a Task Inventory into five major categories called Managerial, Clerical, Heavy, Light, and Dirty tasks. Could we identify a relationship between time-in-service and the type of task being performed?" Since there are no duties with these titles, the five new modules must be defined. As in example 1, we will use a subject matter specialist to identify those tasks which fall under the new module definitions. (Appendix E) Then, four additional categories will be produced representing people who have been in the service 1-24 months, 25-48 months, 49-96 months and more than 96 months. Combining the modules defined earlier with these four descriptions, a report is produced that addresses the user's question. (Appendix F) Another approach might use male/female categories in place of time-in-service.

Example 3: Associating Tasks with Training Standards. The Air Force has established a document for every AFSC called the Specialty Training Standard (STS). Supervisors in the field are familiar with this form and when presenting data to these personnel, it should be organized accordingly. (Appendix G) Again, a subject matter specialist is utilized in associating the STS document with the Task Inventory.

Example 4: Computer Generated Modules. If the requirements from the definition step are quantifiable, then the assignment of tasks to modules can be computer generated. For example, the assignment could be based on the probability of co-performance from a matrix containing the probabilities that tasks are performed together. Using this matrix, tasks which are likely to be performed together cluster into groups called task modules. (Appendix H) The next step is for a subject matter specialist to study those task modules and label each as a separate group such as training module, etc.

Example 5: Relating Tools to Tasks Performed. Suppose we wish to look at the association of tools and equipment with tasks performed. A difference description is produced by comparing job descriptions of those people who do and don't use a selected piece of equipment. This identifies those tasks which are likely to be related to the use of the tool and they become members of the new module. Analysis reports are then generated by merging several tool module descriptions by case membership groups.

SUMMARY

The purpose of any computerized approach to problem solving is to provide the information necessary for making decisions. Computer programs have been developed to produce these decision making reports and the programs take into consideration that the questions asked about the data will differ by application. These programs are now an integral part of the Comprehensive Occupational Data Analysis Programs (CODAP) system at the Air Force Human Resources Laboratory. In conclusion, through the use of user defined modules we have realized a more effective utilization of existing data.

Appendix A. Job Inventory for Vehicle Maintenance

JOB INVENTORY IN ALPHABETICAL SEQUENCE. ONLY 42 OUT OF 690 POSSIBLE TASKS ARE SHOWN.

1. ADJUST VALVE CLEARANCE
2. ADMINISTER OR SCORE TESTS
3. ALIGN OR ADJUST HEADLIGHTS
4. ANALYZE CAUSE OF BRAKE FAILURE
5. ANALYZE CAUSE OF VEHICLE FAILURE
6. ANALYZE MAINTENANCE TRENDS
7. CHANGE ENGINE OIL
8. CHECK OR SERVICE OIL LEVELS
9. CLEAN BATTERY POSTS
10. CONDUCT CLASSROOM TRAINING
11. CONDUCT OR ATTEND STAFF MEETINGS
12. COORDINATE WITH SUPPLIERS TO MAINTAIN REQUIRED PARTS
13. DEMONSTRATE OPERATION OF EQUIPMENT
14. DISASSEMBLE DISTRIBUTORS
15. DRAIN COOLING SYSTEMS
16. DRAFT CORRESPONDENCE
17. ESTABLISH MAINTENANCE PROCEDURES
18. FLUSH TRANSMISSIONS
19. INSPECT BRAKES
20. INSPECT ENGINE VALVE GUIDES
21. INSPECT FRONT END ALIGNMENT
22. INSPECT IGNITION POINTS
23. INSPECT MOTOR MOUNTINGS
24. INSPECT MAINTENANCE RECORDS
25. INSPECT TIRES
26. INSPECT VALVE COVER GASKETS
27. INSTALL BRAKE LININGS
28. INSTALL CYLINDER LINERS
29. INSTALL ENGINES
30. INSTALL POINTS
31. INSTALL TRAILER HITCHES
32. ISSUE OR MAINTAIN STOCK ITEMS OF HIGH VALUE
33. ISSUE PARTS FROM STOCK ROOM
34. MAINTAIN ACCIDENT LOG
35. MAINTAIN INVENTORY FORM 226
36. MAINTAIN VEHICLE MAINTENANCE FORM 100
37. MANUFACTURE ENGINE GASKETS
38. OPERATE ELECTRONIC TEST EQUIPMENT
39. OPERATE TIRE BALANCING EQUIPMENT
40. PLAN AIDS FOR TRAINING
41. PREPARE ACCIDENT REPORT FORM 22
42. PREPARE BRIEFINGS

JOB INVENTORY IS CATEGORIZED BY DUTIES WITH THE APPLICABLE TASKS SHOWN IN ALPHABETICAL SEQUENCE. SOME DUTIES AND TASKS ARE NOT SHOWN.

A. ORGANIZING, PLANNING AND MANAGING
  6. ANALYZE MAINTENANCE TRENDS
  11. CONDUCT OR ATTEND STAFF MEETINGS
  16. DRAFT CORRESPONDENCE
  17. ESTABLISH MAINTENANCE PROCEDURES
  24. INSPECT MAINTENANCE RECORDS
  42. PREPARE BRIEFINGS

B. TRAINING
  2. ADMINISTER OR SCORE TESTS
  10. CONDUCT CLASSROOM TRAINING
  13. DEMONSTRATE OPERATION OF EQUIPMENT
  40. PLAN AIDS FOR TRAINING
  44. PREPARE LESSON PLANS
  53. SELECT INDIVIDUALS TO ATTEND TRAINING

C. WORKING WITH FORMS
  35. MAINTAIN INVENTORY FORM 226
  36. MAINTAIN VEHICLE MAINTENANCE FORM 100
  41. PREPARE ACCIDENT REPORT FORM 22
  45. PREPARE SURPLUS INVENTORY FORM 695-7

D. PERFORMING SUPPLY FUNCTIONS

  12. COORDINATE WITH SUPPLIERS TO MAINTAIN REQUIRED PARTS
  32. ISSUE OR MAINTAIN STOCK ITEMS OF HIGH VALUE
  33. ISSUE PARTS FROM STOCK ROOM
  51. RESEARCH FEDERAL STOCK NUMBERS OR PART NUMBERS
  54. STOCK PARTS, SUPPLIES OR EQUIPMENT

E. TROUBLESHOOTING VEHICLES
  4. ANALYZE CAUSE OF ENGINE FAILURE
  5. ANALYZE CAUSE OF BRAKE FAILURE
  20. INSPECT ENGINE VALVE GUIDES
  22. INSPECT IGNITION POINTS
  23. INSPECT MOTOR MOUNTINGS
  52. ROAD TEST VEHICLES

F. REMOVING, REPLACING OR CLEANING PARTS
  7. CHANGE ENGINE OIL
  9. CLEAN BATTERY POSTS
  15. DRAIN COOLING SYSTEMS
  19. INSTALL BRAKE LININGS
  28. INSTALL CYLINDER LININGS
  30. INSTALL POINTS
  48. REMOVE OR REPLACE PISTONS AND RINGS

Appendix C. Job Inventory for Vehicle Maintenance

JOB INVENTORY IS IN ALPHABETICAL ORDER WITH ASTERISKS PLACED BY THOSE TASKS IDENTIFIED BY A SUBJECT MATTER SPECIALIST AS PART OF OVERHAULING AN ENGINE.

*1. ADJUST VALVE CLEARANCE
2. ADMINISTER OR SCORE TESTS
3. ALIGN OR ADJUST HEADLIGHTS
4. ANALYZE CAUSE OF BRAKE FAILURE
5. ANALYZE CAUSE OF VEHICLE FAILURE
6. ANALYZE MAINTENANCE TRENDS
7. CHANGE ENGINE OIL
*8. CHECK OR SERVICE OIL LEVELS
9. CLEAN BATTERY POSTS
10. CONDUCT CLASSROOM TRAINING
11. CONDUCT OR ATTEND STAFF MEETINGS
12. COORDINATE WITH SUPPLIERS TO MAINTAIN REQUIRED PARTS
13. DEMONSTRATE OPERATION OF EQUIPMENT
*14. DISASSEMBLE DISTRIBUTORS
*15. DRAIN COOLING SYSTEMS
16. DRAFT CORRESPONDENCE
17. ESTABLISH MAINTENANCE PROCEDURES
18. FLUSH TRANSMISSIONS
19. INSPECT BRAKES
*20. INSPECT ENGINE VALVE GUIDES
21. INSPECT FRONT END ALIGNMENT
*22. INSPECT IGNITION POINTS
23. INSPECT MOTOR MOUNTINGS
24. INSPECT MAINTENANCE RECORDS
25. INSPECT TIRES
*26. INSPECT VALVE COVER GASKETS
27. INSTALL BRAKE LININGS
*28. INSTALL CYLINDER LINERS
*29. INSTALL ENGINES
30. INSTALL POINTS
31. INSTALL TRAILER HITCHES
32. ISSUE OR MAINTAIN STOCK ITEMS OF HIGH VALUE
33. ISSUE PARTS FROM STOCK ROOM
34. MAINTAIN ACCIDENT LOG
35. MAINTAIN INVENTORY FORM 226
36. MAINTAIN VEHICLE MAINTENANCE FORM 100
37. MANUFACTURE ENGINE GASKETS
*38. OPERATE ELECTRONIC TEST EQUIPMENT
39. OPERATE TIRE BALANCING EQUIPMENT
40. PLAN AIDS FOR TRAINING
41. PREPARE ACCIDENT REPORT FORM 22
42. PREPARE BRIEFINGS

THIS MODULE SHOWS THOSE TASKS IDENTIFIED AS APPLICABLE TO OVERHAULING ENGINES. SOME TASKS NOT SHOWN.

		PERCENT MEMBERS PERFORMING	
		1ST TERM	2ND TERM
A.	OVERHAULING ENGINES		
1.	ADJUST VALVE CLEARANCE	2.0	32.1
8.	CHECK OR SERVICE OIL LEVELS	48.7	4.6
14.	DISASSEMBLE DISTRIBUTORS	4.7	62.7
15.	DRAIN COOLING SYSTEMS	36.2	10.5
20.	INSPECT ENGINE VALVE GUIDES	1.1	38.7
22.	INSPECT IGNITION POINTS	4.8	56.3
26.	INSPECT ENGINE COVER GASKETS	25.9	19.2
28.	INSTALL CYLINDER LINERS	0.5	66.9
29.	INSTALL ENGINE	30.3	26.4
38.	OPERATE ELECTRONIC TEST EQUIPMENT	1.1	43.6
46.	PREPARE VEHICLE MAINTENANCE FORM 100	57.4	10.1
48.	REMOVE OR REPLACE PISTONS OR RINGS	5.3	26.7
49.	REMOVE OR REPLACE POINTS	2.4	47.8

401

Appendix E. Five Categorical Modules

THESE MODULES SHOW WHICH TASKS WERE IDENTIFIED AS BELONGING TO THE SPECIFIED CATEGORY. SOME TASKS NOT SHOWN.

A. MANAGERIAL

6. ANALYZE MAINTENANCE TRENDS
11. CONDUCT OR ATTEND STAFF MEETINGS
16. DRAFT CORRESPONDENCE
24. INSPECT MAINTENANCE RECORDS
42. PREPARE BRIEFINGS

B. CLERICAL

2. ADMINISTER OR SCORE TESTS
16. DRAFT CORRESPONDENCE
34. MAINTAIN ACCIDENT LOG
35. MAINTAIN INVENTORY FORM 226
46. MAINTAIN VEHICLE MAINTENANCE FORM 100

C. HEAVY TASKS

29. INSTALL ENGINES
33. ISSUE PARTS FROM STOCK ROOM
47. REMOVE OR REPLACE BATTERIES
50. REMOVE OR REPLACE POWER STEERING UNITS
55. ROTATE TIRES

D. DIRTY TASKS

7. CHANGE ENGINE OIL
9. CLEAN BATTERY POSTS
15. DRAIN COOLING SYSTEMS
29. INSTALL ENGINES
47. REMOVE OR REPLACE BATTERIES

E. LIGHT TASKS

1. ADJUST VALVE CLEARANCES
3. ALIGN OR ADJUST HEADLIGHTS
7. CHANGE ENGINE OIL
9. CLEAN BATTERY POSTS
49. REMOVE OR REPLACE POINTS

Appendix F. Percent Members Performing Categorical Modules

THE MODULES, WITH RELATED TASKS, SHOW WHICH CATEGORY OF PEOPLE ARE PERFORMING WHAT TYPE OF TASK. PERCENT MEMBERS PERFORMING DATA IS USED.

	1 -24	25 -48	49 -96	96+
A. MANGERIAL				
6. ANALYZE MAINTENANCE TRENDS	0.0	0.7	3.1	25.2
11. CONDUCT OR ATTEND STAFF MEETINGS	1.1	1.7	10.7	89.3
16. DRAFT CORRESPONDENCE	4.3	10.2	36.7	5.5
24. INSPECT MAINTENANCE RECORDS	1.6	5.6	42.3	10.5
42. PREPARE BRIEFINGS	0.0	0.0	15.6	75.6
B. CLERICAL				
2. ADMINISTER OR SCORE TESTS	5.6	32.1	10.5	1.1
16. DRAFT CORRESPONDENCE	4.3	10.2	36.7	5.5
34. MAINTAIN ACCIDENT LOG	4.7	50.1	48.6	5.9
35. MAINTAIN INVENTORY FORM 226	16.3	42.8	10.2	0.5
46. MAINTAIN VEHICLE MAINTENANCE FORM 100	10.2	66.6	12.7	0.1
C. HEAVY TASKS				
29. INSTALL ENGINES	26.5	30.3	8.8	0.9
33. ISSUE PARTS FROM STOCK ROOM	10.9	26.3	25.5	1.3
47. REMOVE OR REPLACE BATTERIES	63.7	10.2	1.6	0.0
50. REMOVE OR REPLACE POWER STEERING UNITS	15.1	26.9	9.9	2.6
55. ROTATE TIRES	72.6	21.0	4.1	0.0
D. DIRTY TASKS				
7. CHANGE ENGINE OIL	66.7	30.9	4.1	0.0
9. CLEAN BATTERY POSTS	54.3	5.0	1.1	0.0
15. DRAIN COOLING SYSTEMS	51.6	4.7	2.6	0.1
29. INSTALL ENGINES	25.6	30.3	8.8	0.9
47. REMOVE OR REPLACE BATTERIES	63.7	10.2	1.6	0.0
E. LIGHT TASKS				
1. ADJUST VALVE CLEARANCES	0.0	5.1	32.6	15.5
3. ALIGN OR ADJUST HEADLIGHTS	20.1	22.6	5.4	1.0
7. CHANGE ENGINE OIL	66.7	30.9	4.1	0.0
9. CLEAN BATTERY POSTS	54.3	5.0	1.1	0.0
49. REMOVE OR REPLACE POINTS	2.9	18.7	26.1	3.0

TASKS ARE ASSOCIATED WITH THE SPECIALTY TRAINING STANDARD FOR THE VEHICLE MAINTENANCE PERSONNEL.

IA DISASTER PREPAREDNESS & EMERGENCY PROCEDURES

151. ATTEND SAFETY BRIEFINGS
102. MAINTAIN FIRE EXTINGUISHER READINESS FORM 672
133. PERFORM SPOT CHECKS OF SAFETY READINESS
264. PRACTICE EMERGENCY PROCEDURES

IIB SECURITY

32. ISSUE OR MAINTAIN STOCK ITEMS OF HIGH VALUE
196. MAINTAIN STOCK INVENTORY
599. PLAN SECURITY PROGRAMS
602. CONDUCT SECURITY BRIEFINGS

IVA SUPERVISING AND TRAINING

2. ADMINISTER OR SCORE TESTS
10. CONDUCT CLASSROOM TRAINING
220. SCHEDULE WORK ASSIGNMENTS
319. SUPERVISE SUBORDINATES

Appendix H. Computer Generated Module

THESE MODULES WERE GROUPED TOGETHER BASED ON THEIR PROBABILITY OF BEING PERFORMED TOGETHER.

A. MINOR ENGINE OR TRANSMISSION SERVICING

7. CHANGE ENGINE OIL
8. CHECK OR SERVICE OIL LEVELS
9. CLEAN BATTERY POSTS
15. DRAIN COOLING SYSTEMS
18. FLUSH TRANSMISSIONS

B. SERVICING ELECTRICAL SYSTEMS

14. DISASSEMBLE DISTRIBUTORS
22. INSPECT IGNITION POINTS
38. OPERATE ELECTRONIC TEST EQUIPMENT
49. REMOVE OR REPLACE POINTS

C. CLASSROOM TRAINING

2. ADMINISTER OR SCORE TESTS
10. CONDUCT CLASSROOM TRAINING
13. DEMONSTRATE OPERATION OF EQUIPMENT
15[illegible]. OPERATE AUDIOVISUAL EQUIPMENT
170. PLAN AIDS FOR TRAINING
43. PREPARE LESSON PLANS
189. SIGN OFF TRAINING RECORDS

OCCUPATIONAL ANALYSIS FOR FIELD GRADE ARMY OFFICERS

Sally J. Van Nostrand

US Army Research Institute for the Behavioral and Social Sciences

and

M. Reid Wallis

Richard A. Gibboney Associates, Inc.

ABSTRACT

The Command and General Staff College (CGSC) prepares Army officers for duty as field grade commanders and principal staff officers at brigade and higher echelons. The College consumes significant expenditures and provides the first, and for the majority of field grade officers the only formal Army training for high level jobs. Despite the importance of the CGSC mission, occupational definition of post-CGSC assignments and the crosswalks to training needs analysis at this level of responsibility have not yet been objectively addressed. In a memorandum to the Army Research Institute (ARI) in 1977 the CGSC Commandant stated, "front-end analysis to support curriculum development . . . is one of the most pressing priorities that the College faces today." He requested that ARI research the feasibility of using the ARI Duty Module concept "to provide an information base for decision on further research effort and its direction."

This research was directed to the examination of two disparate sub-courses of the CGSC curriculum. Research design, results from the feasibility prototype, and directions for further research are discussed.

BACKGROUND

ARMY RESEARCH INSTITUTE INVOLVEMENT WITH COMMAND AND GENERAL STAFF COLLEGE CURRICULUM DEVELOPMENT

In a memorandum to the Army Research Institute (ARI) in 1977 the Command and General Staff College (CGSC) Commandant stated, "front-end analysis to support curriculum development ... is one of the most pressing priorities that the College faces today." He requested that ARI research the feasibility of using the ARI Duty Module methodology "to provide an information base for decision on further research effort and its direction." The feasibility research has been completed. ARI is currently working in both the Analysis and Control (external evaluation or feedback) phases of the Instructional Systems Development (ISD) of CGSC curriculum development. The ongoing research was precipitated by the Duty Module feasibility results, statements of Human Resource Needs (HRNs) for new methods of front-end analysis for non-procedural tasks from several Army schools and HRNs for feedback on training and education[1] from CGSC graduates.

CGSC MISSION REQUIRES BOTH TRAINING AND EDUCATION

The mission of the Command and General Staff College[2] is to provide instruction for officers of the Active Army and Reserve components, worldwide, so as to prepare them for duty as field grade commanders and principal staff officers at brigade and higher echelons.

The College prepares officers to:

-- Command battalions, brigades, and equivalent-sized units in peace or war.

-- Train these units to accomplish their assigned missions.

-- Employ and sustain weapon systems to optimize their effect in the conduct of combined arms operations.

-- Serve as principal staff officers from brigade through division, to include support commands, and as staff officers of higher echelons, including major Army, joint, unified, or combined headquarters.

---

[1] The definitions of training and education for this paper are: Training - Teaching specific skills which will be needed in the next assignment. Education - Teaching broad knowledge areas as a foundation for the requirements of all expected positions in the future, not necessarily for the next assignment.

[2] 1977-78 Catalogue, US Army Command and General Staff College

CGSC offers a Master's degree in Military Arts and Sciences, and offers the opportunity to obtain numerous other Master's level degrees from a number of other colleges and universities. Although the junior officer schools (Basic Course for second lieutenants and Advanced Course for captains) teach some basic management and supervisory skills, the major emphasis is on specialty-related tasks and separate schools are run by the specialty branches--the graduate of a Basic or Advanced School is expected to be technically proficient in specialty skills.

In a survey of general officers concerning the Army officer education and training programs (Van Nostrand and Wallis, 1978) attitudes were identified as follows:

-- Management should be taught (at CGSC) but not at the Basic and Advanced Courses where officers are taught to be technicians in their branch specialties.

-- CGSC should teach those brilliant young officers who are to provide the staff and general officers who will run the Army for the next 10 to 20 years. (Approximately 6-7 years after attending CGSC the officers are competitively selected to attend the Army War College)

-- The Army should go to the university concept.

-- What should be taught:

Conceptualization, even though difficult

Develop truly general staff officers

Research, write and brief on solutions to real issues

ISD USED FOR CURRICULUM DEVELOPMENT

All of the US Army schools for officers, except the US Military Academy and the Army War College, are monitored by the US Army Training and Doctrine Command (TRADOC). Curriculum development within TRADOC doctrine requires that the TRADOC monitored schools use the Instructional Systems Development (ISD) process as a systems approach to the development and evaluation of training (TRADOC Pamphlet 350-30). Although ARI research is concerned with all five phases (Analyze, Design, Develop, Implement and Control) of the ISD model, this paper is directed to those phases which require occupational analysis to provide decision-making data. These are:

Analyze - (a) Determine tasks to be taught (front-end analysis)

(b) Determine setting in which each task will be taught

Control - (a) Internal evaluation -- how well did students meet the stated objectives?

(b) External evaluation -- how well do graduates perform on the job? Usually determined by performance evaluations of the school-taught tasks with feedback information to the schools.

Procedures evolved through the use of ISD in the Army schools have proved useful; they represent many person-years of effort to develop a workable, systematized training approach. Some of these procedures are:

a. Occupational description techniques developed to define a position in terms of tasks having specific beginning and ending times, cue to perform, and step-by-step (or procedural) description of how the task is to be performed. These techniques have proved useful for the majority of enlisted tasks and for many of the NCO and company grade officer specialty-unique tasks. The majority of the Army schools responsible for training for these jobs need concern themselves with only those jobs which are unique to their specialties.

b. The crosswalk from occupational analysis to training requirements has been successfully addressed for enlisted personnel. However, the problem of training requirements of supervisors and managers at the non-commissioned officer (NCO) level based on job descriptions has not yet been resolved. This problem has already surfaced for company grade officers in the recently initiated TRADOC program for defining officer tasks.

c. Criticality has been refined to four measures commonly called "the four-factor model." However, this refinement is inadequate to answer all criticality questions.

d. A concept that permeates all descriptions of ISD is, "train for the next job to be performed" i.e. if the trainee will not use the skill very soon there may be no reason for training it -- the learning retention decay rate may prove the training resources could better be allocated elsewhere.

The ISD process is proving to be very difficult to implement at the Command and General Staff College. The standards or concepts noted above, although not necessarily "standard" in the original ISD reports (Branson, et al, 1975) are particularly difficult to apply to the curriculum for field grade officers. The CGSC curriculum which does not focus on specialty proficiency, but rather a general broadening of horizons for field grade officers, cannot be fitted to the conventional front-end analysis techniques of the ISD process.

First, as CGSC serves the entire Army, not just a few specialities the sheer size of the data base is a problem--all field grade officer positions in the Army must be subjected to occupational analysis for creation of the task lists. Using the assumption (as is usually now the case) that a supervisor's job must include generalized management tasks plus a knowledge of the tasks of all the supervised personnel, the size of the data base is multiplied by some unknown factor.

Second, a unit of instruction usually teaches several related tasks. As the data base becomes larger it becomes more and more difficult to find all of the related tasks. Unfortunately, the task analysis techniques do not yield tasks which fit clustering requirements for CGSC curriculum development.

Next, CGSC is a masters level degree granting institution, and is in this respect, unique among the TRADOC schools. The concept of CGSC as an institute of higher learning, providing the foundation for future, individual officer self-development and growth (to "think and decide") requires that subjects be taught which are not based on "next assignment," but are general education in many different fields.

Further, as CGSC is the formal training/education institution for the Army "middle managers," many of the tasks for which CGSC does train are non-procedural in nature, i.e., these tasks are difficult, perhaps impossible to define in terms of cue to perform, begin and end points, steps to perform, and evaluation criteria.

Even more difficult is the choosing of criteria on which to base the train/don't train decision. The four factor criteria used for enlisted and branch specific tasks do not apply. A concept that has been popular recently is, "the officer is much more than the sum of those skills in which proficiency can be demonstrated." Consider the following hypothetical example: If most field grade officers spend 50% of their time reading paper work of some type and less than 1% of the time making decisions; should CGSC train them to read paper work, or should more resources be spent in teaching good decision-making?

PREVIOUS FIELD GRADE OFFICER OCCUPATIONAL ANALYSIS RESEARCH BY ARI

Responding to personnel management needs ARI has been working on the Duty Module concept since 1970. A Duty Module represents a significant work activity; is applicable to a number of different duty positions, and describes the various jobs in a common language. A Duty Module is smaller than an MOS or any one job within an MOS and larger than a task. It is actually a cluster of 10 to 20 tasks that relate, occupationally and organizationally, in meaningful ways. These tasks are very much like the tasks produced by other job analysis techniques, but the significant difference is that a major emphasis of the original research was to produce meaningful task clusters. These horizontal clusters can be used as building blocks, or "plug-in" units, to describe the significant duties of any job using only a few Duty Modules. Duty Modules are also designed for describing jobs at all levels of responsibility (vertically clustered). Therefore, the full interrelationship among jobs, across all specialties and for all officer grades, both similarities (commonality) and differences, can be codified.

Although the Duty Module methodology could be applied to civilian organizations, or to enlisted or NCO positions, the research was directed to support of the Officer Personnel Management System (OPMS) and the present data base is essentially complete for officer duties common to all the OPMS specialties. Further development to complete the OPMS data base would necessitate creation of only a limited number of specialty-specific Duty Modules.

APPLICATION OF DUTY MODULE TECHNIQUE TO FRONT-END ANALYSIS

Most CGSC graduates will be assigned as a staff officer, some at very high levels, others may assume command of a battalion or brigade. The commander's management role is analagous to that of the operations manager of a medium-sized manufacturing company. Additional duties of the position require responsibility for the unit as it trains to achieve and maintain combat readiness during peacetime, with the capability for rapid transition to combat effectiveness during war. The resources available to, and, therefore, controlled by, one Armor battalion commander consists of approximately 550 personnel, $55 million investment in equipment, and annual expenditures of $13 million. The staff role of the CGSC graduate can have comparable responsibility. In context of the increasingly constrained training resources, the growing importance of training quality can not be overstressed. As the quality of training is dependent upon the adequacy of the front-end analysis, those responsible for CGSC curriculum development have a continuing concern with development of better front-end techniques. In keeping with this concern, the CGSC has used both formal and informal channels to obtain feedback on the appropriateness and utility of the instruction. This concern has stimulated many students to study some aspects of curriculum development as part of their independent research requirement.

A recently completed CGSC Master's thesis (Norris and Robbins, 1977) explored the feasibility of utilizing Duty Modules for the front-end analysis of the CGSC regular course. The thesis is based in part upon earlier ARI work, Cory, Medland, and Uhlaner (1977); Davis and

411

Korotkin (1975); Korotkin, et.al. (1975); and others. This thesis develops the concept of Duty Modules as the vehicle for the ISD Analysis phases of CGSC curriculum development.

Although Norris and Robbins point out some possible shortcomings of the Duty Module approach, they nontheless conclude that theoretically,

> "....Duty Modules offer an attractive approach to this problem and have the major advantage of being beyond the 'drawing board stage'. Duty Modules are a reality and the effort in time and resources to apply these concepts to the college is far less than that required to develop new methodology."

The need for empirical validation of the Norris and Robbins approach stimulated the CGSC Commandant's request that ARI conduct the prototype feasibility research which was initiated during the fall of 1977. The design of the prototype analysis was:

a. Identify two significant assignments filled by CGSC graudates.

b. Identify the CGSC courses or sub-courses which prepared the officer for the identified assignments.

c. Describe both the course curriculum and the assignments using the Duty Module structure.

d. Compare each assignment Duty Module structure with the Duty Module structure of the related CGSC course. Commonality will be indicative of degree of correlation between training and job requirements. Significant commonality would indicate a high degree of overlap between content taught and skills required on the job. Lack of or little commonality would indicate one or more of the following:

1. CGSC is teaching material not required or necessary to the job.

2. CGSC is not teaching skills required by the assignment.

3. The Duty Module approach is not feasible.

Two assignment areas were selected to represent disparate duties and relate to specific instructional areas:

a. Combat commander; related course is "Battle Captains"

b. Staff assignments at Office of the Secretary of Defense (OSD), Office of the Joint Chiefs of Staff (OJCS), Department of the Army (DA), and Army major commands; related course is "High Level Staff Applications."

The Battle Captains course is one course of a sequence of five orientation courses given as refresher training to command designees (lieutenant colonel and colonel) prior to their assumption of command. Each of these five courses closely matches one of the six previously validated Duty Modules which apply to unit commanders, although detailed analysis was not performed for the four not taught at CGSC. The subject matter of one of the Duty Modules, "General Administration," is not taught in these orientation courses and it must be assumed that the officer retains the necessary knowledge and skills from previous education and on-the-job training.

Comparison of the detailed task analyses of the Battle Captains course and of the O-U-1 Duty Module, "Directs and controls employment of Infantry and Armor maneuver unit," shows that the tasks taught and the tasks performed correspond exactly. Using the same technique it should be possible to compare the other four orientation courses and, if necessary, to develop another couse for the general administration module. For this course we can say that the Duty Module front-end analysis procedure is feasible.

The comparison of relevant duty modules and the High Level Staff Application Course was more difficult. To adequately describe the position, "Action Officer, High Level," it was necessary to create one new Duty Module, "Performs action officer functions on a high level staff." Verification of this new Duty Module was accomplished by interviewing a sample of 20 respondents holding high level staff positions. Although all 20 respondents performed the new module, it was necessary to use 17 Duty Modules from the data base to adequately describe their positions. It is unusual to need as many as 18 for 20 similar positions, but the job incumbents represented 8 different branches, 11 primary specialties and 12 alternate specialties (a total of 19 different specialties). The 18 Duty Modules performed by the surveyed incumbents were all, except the new one, specialty related and, therefore, would not be of concern to CGSC; they would, or should have been taught at the specialty related schools and earlier attendance at CGSC. These modules had been verified in earlier research but were, however, examined to assure that they did continue to accurately describe the duties.

An examination of the program of instruction (POI) revealed these five subject areas:

a. The organization, functions and relationships between OSD, OJCS, Office of the Secretary of the Army (OSA), and Office of the Chief of Staff of the Army (OCSA).

b. The organization, functions and relationship to DA of

- Headquarters, TRADOC

- Headquarters, DARCOM

- Headquarters, FORSCOM,

c. The organization, functions and relationship of Headquarters, US Readiness Command to OJCS.

d. DA staffing procedures to include rewriting a decision memorandum into 175 words or less and writing two information papers of 175 words or less.

e. Staff techniques and procedures used within the OJCS. Of the five subject areas, the last two listed, being performance oriented, lend themselves to a front-end analysis using Duty Module techniques.

The first three subjects are informational in nature and cannot be directly translated into a Duty Module structure.

HIGH LEVEL STAFF COMPARISON

CGSC COURSE: HIGH LEVEL STAFF	DUTY MODULE: PERFORMS ACTION OFFICER FUNCTIONS ON A HIGH LEVEL STAFF
SUBJECT AREAS:	TASKS:
a. Organization, functions and relationships between OSD, OJCS, OSA AND OCSA.	a. Prepare decision memoranda, information memoranda, information papers, and other similar documents for a superior.
b. Organization, functions and relationships to DA of TRADOC, DARCOM and FORSCOM.	b. Represent superior in action officer meetings.
c. Organization, functions and relationships of US Readiness Command to OJCS.	c. Process joint staff action directives.
d. DA staffing procedures including writing decision memorandum and two information papers.	
e. Staff techniques and procedures used within the OJCS.	

FIGURE 1

Comparing subject area d and e from the curriculum with the tasks in the new Duty Module, one can see a close correlation, see Figure 1. This signifies that these subject areas should be included in the course curriculum. This type of comparison, however, does not lend itself to a statistical analysis so it is not possible to state a confidence level with which one can say they should be included, or what percentage of the time should be devoted to them, especially as only some, not all of the officers use the OJCS staffing procedures.

To explore the applicability of the methodology to the first three subject areas a questionnaire was administered. Respondents were asked to indicate the degree of understanding, ranging from "comprehensive" to "no understanding," which they needed of OSD, OJCS, DA TRADOC, FORSCOM, DARCOM, US Readiness Command, or other similar headquarters in order to perform their assigned duties. Not surprisingly, the survey sample composed of DA and DARCOM staff officers indicated a need for a high level of understanding of the organization and functions of their own headquarters. Next followed OSD, TRADOC, FORSCOM, OJCS, and US Readiness Command, in that order. One can deduce that the "need to know" rating of any headquarters would go up if officers from that headquarters were included in the survey sample. It does appear significant, however, that the US Readiness Command received lower need to know ratings from the survey sample than did write-ins for US Army Europe (USAREUR). This outcome suggests that consideration be given to examining whether Headquarters, USAREUR should replace US Readiness Command in the POI. Before this consideration, however, a larger survey which includes officers from all of the designated offices should be performed. If the result still holds true the POI decision should be made by training experts; there may be valid reasons for including a joint headquarters in the curriculum to the exclusion of a major overseas command.

When courses teach performance-oriented skills, it is logical that the skills should appear in a Duty Module for some Army job, as Duty Modules are a prior performance-oriented. When courses teach information, that information will not appear in a Duty Module directly, but only in a performance task which is influenced by the information acquired. This is easily seen by the results of the two comparisons. The skills taught by the Battle Captains Course are performance-oriented; the Duty Module approach was completely successful for a front-end analysis of this course. The High-Level Staff Application Course teaches some performance skills and some knowledges (information); the Duty Module approach was only partially successful for this front-end analysis.

RESULTS

The results of the feasibility can be stated:

a. For performance-oriented skills it is feasible to use Duty Modules to fully describe all positions filled by graduates of CGSC, then compare the applicable Duty Modules with the Duty Module structures

415

of the scope and instructional objectives of a large portion of the college curriculum to "identify curriculum needs and define CGSC output, both critical elements in resource justification" (Norris and Robbins, Ibid).

b. The Duty Module approach is not adequate for those courses or sub-objectives of courses which are designed to impart knowledges or information; expanded front-end analysis techniques must be developed for these.

RESEARCH IN PROGRESS

FRONT-END ANALYSIS FOR NON-PROCEDURAL TASKS

Several of the schools in the TRADOC community have identified needs for new front-end analysis techniques for those parts of the curriculum which are difficult to describe in terms of performance-oriented (procedural) tasks such as administrative, communication skills (interpersonal as well as reading, writing and briefing), and leadership. This first effort is to make more explicit the procedures for defining these non-procedural assignment requirements which should be included in education/training programs but which are not normally described in officer job descriptions. Several alternative methods for representing these additional data for inclusion in job analyses are being considered. Some of these are:

a. A simple task list prepared in CODAP-type format using the scales developed for enlisted positions,

b. A similar CODAP-type format, but using scales which "expert opinion" feels should be used for officer surveys,

c. A questionnaire at the level of topics in a Program of Instruction rather than tasks (this alternative will also examine alternative types of responses such as simple "yes" or "no" responses to the question, "is it needed?", to having a job incumbent allocate proportionate times of instructional hours that s/he feels would optimally prepare someone for an assignment.), and

d. A questionnaire on the POI but using the CODAP task format.

As this is an exploratory effort the questions that we hope will be answered are qualitative, not quantitative. These are:

a. Can these skills be adequately described in terms of tasks and/or topics?

b. Are the representations of these skills meaningful to job incumbents who have been trained in these skills and are now in positions where these skills are required?

c. Are the data meaningful to CGSC curriculum development personnel?

TRAINING INFORMATION FEEDBACK SYSTEM

Concurrent with the exploratory effort for non-procedural tasks is the development of a Training Information Feedback System (TIFS) for CGSC. The objective is to create clusters of similar tasks into data elements called Job Certification Components (JCCs) which can be used in computerized data bases for individual officer competency certification, for feedback to curriculum developers, career management, and specialty proponents professional development programs.

The JCCs can be useful for curriculum development if it is possible to show differential performance between those officers who have attended the appropriate education/training course(s) and officers who have not attended but are nonetheless serving in the same positions as graduates. The target population, therefore, is made up of field grade officers serving in the same positions, the first set being graduates of CGSC and the second set being those who have not attended. Occupational analysis techniques will be used for development of JCCs for these duty positions. JCCs will then be verified as useful for a TIFS for CGSC by analysis of job performance of officers from both sets.

A Technique for Selecting
Electronic Specialties for Consolidation

by

Hendrick W. Ruck

Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper are those of the author and are not necessarily those of the United States Air Force.

The Air Force occupational classification system for enlisted personnel is composed of approximately 250 specialties. These specialties cover a wide range of occupations such as band members, medical technologists, pneudraulics repairmen, and aircraft control and warning radar repairmen. Approximately 50 of these specialties are generally considered to be "electronic specialties." These electronic specialties (see Table 1) are vital to the Air Force, since the airmen in these specialties have the responsibility for maintaining the Air Force's global communications network, defensive surveillance systems, and air navigation and communication systems. Airmen in these specialties comprise somewhat more than 10 percent of the enlisted force. Even more important, though, is the investment the Air Force makes in training these airmen. Technical training designed to give initial skills to airmen in electronic specialties is costlier than technical training in other specialties in both time and equipment. For purposes of efficient personnel management, effective personnel utilization, and efficacious training, the Air Force is seriously considering consolidating several of these specialties.

Table 1
Examples of "Electronic" Specialties

302X0	Weather Equipment Specialist
305X4	Electronic Computer Systems Specialist
316X2	Missile Electronic Equipment Specialist
316X3	Instrumentation Mechanic
321X1	Defensive Fire Control Systems Mechanic
328X3	Electronic Warfare Systems Specialist
341X3	Analog Flight Simulator Specialist
341X6	Digital Navigation/Tactics Training Devices Specialist
361X0	Outside Wire & Antenna Maintenance Repairman
362X3	Missile Control Communications Systems Specialist
403X0	Biomedical Equipment Maintenance Specialist

Consolidation of specialties offers the Air Force several advantages. These advantages include (a) increased operational flexibility in utilization of personnel within field units, (b) simpler assignments due to larger pools of eligible incumbents and fewer specialties, (c) simpler training due to fewer initial-skill courses, and (d) reduced manning, since specialists would have broader expertise and therefore fewer specialties (and specialists) would be involved in maintaining complex systems.

Electronic principles are relatively well defined and are generally regarded as necessary prerequisite knowledge for job proficiency in electronic specialties. The assumption that there are underlying principles that are common across electronic specialties has offered the possibility of studying the actual overlap in electronic principle utilization among these specialties. Electronic principles are rather easily identified, since the Air Force offers common core courses in electronics as prerequisites to entry into the equipment portion of specialist courses. The purpose of this paper is to present preliminary results of a commonality analysis among 20 of the electronic specialties and to discuss the procedures used in the analysis. The electronic specialties analyzed in this study are all from two career fields, Communications-Electronics Systems and Wire Communications Systems. These fields contain 24 specialties that maintain ground communications sytems. Air Force managers have expressed interest in reducing the number of ground electronics specialties for reasons described earlier.

Commonality and Consolidation Considerations

When looking at the feasibility of consolidating specialties, information concerning at least three personnel-related subsystems is required: the training, manning, and recruiting subsystems. An outline of the information that must be synthesized and analyzed in the process of making consolidation decisions is presented in Table 2.

Table 2
Some Considerations Relating to Consolidation of Specialties

Training	Manning	Recruiting
Equipment Similarity	Work Center Location	Recruiting Difficulty
Job/Task Similarity	Total Manning	Aptitude Requirements
Underlying Principles/ Knowledge Similarity	CONUS/Overseas Ratio	Attrition
	Unit Manning	

Data relating to the similarity of equipment maintained or used in different specialties can be gathered from routine occupational surveys, special surveys, logistics or functional managers, or technical orders. Regardless of the data source, however, expert judges would be required to provide measures of similarity. In the ground communications electronics area alone (24 specialties) over 1400 end items are managed within the Air Force logistics system. It is difficult to estimate how many additional major command specific end items are in the inventory. If one extends a similarity analysis to the total electronics community, it can be readily seen that generating similarity overlap measures on end items of equipment for electronic specialties would be an overwhelming task. The difficulty in such an analysis is that judges familiar with several items of equipment would be required to estimate similarity. Aside from the statistical problem of combining judgments made on different combinations of equipment end items, there are at least two other difficulties. One is the number of judges that may be required, and the other is the definition of the dimensions of similarity.

As a second approach, one might ask, then, how difficult is it to gauge the similarity of jobs and tasks performed in the electronics community? A ready source of data exists since Air Force occupational survey data have been collected and analyzed for most of the specialties in question. Once again, the scope of the problem limits the appropriateness of this approach. Several thousand tasks performed by persons in several hundred job types have been identified in the occupational analysis of electronic specialties. Even where tasks are worded similarly, experts must judge equivalence between tasks. Since no acceptable taxonomy has been developed for tasks or job types, it would be an extremely laborious task to compare all tasks with one another or all job types with one another.

What, then, could be done to reduce the magnitude of the comparison problem? Developing a priori groupings of specialties that are good candidates for consolidation is one solution. Various groupings have been developed and sponsored by different Air Force and major command managers. Unfortunately, these grouping schemes have rarely been consistent with one another and, therefore, considerable debate has arisen concerning each proposal. An empirical approach to developing grouping schemes would be possible since data on utilization of fundamental electronic principles could be available for all of the specialties in question. The rationale for this approach would be that initial groupings based on common underlying principles should be developed and subsequent analyses may then be performed for specialties with high commonality in principles. Under this approach the assumption is made that it would be unwise to consolidate specialties that have little or no commonality in principles or knowledge used on the job. Although this assumption may appear at first to be trite, it may be appreciated by those who have familiarity with some of the decisions on specialty structure that have been made in the past.

The considerations relating to manning are somewhat easier to measure than the training considerations. At the unit level, certain specialties have been traditionally undermanned, while others receive priority manning. For the problem at hand, ground communications-electronics maintenance, unit manning is critical. This is due to the fact that many of the positions require round-the-clock manning by fully-qualified personnel. Units that require more than one specialty on 24-hour duty should be identified, and, if there is agreement, those specialties would be good candidates for consolidation providing enough job similarity exists. CONUS/overseas ratio considerations are easily measured in terms of ratios. Traditionally, specialties with high overseas imbalances have been suggested as candidates for merging with specialties that have high CONUS ratios. Again, such possibilities should be tempered with job similarity measures.

Recruiting considerations are more difficult to use in making consolidation decisions. Obviously, specialties that are consolidated should have similar aptitude requirements. However, the impact of recruiting difficulty and attrition on consolidation decisions requires policy makers' and researchers' attention. It is not clear whether it would be in the best interest of the Air Force to merge specialties with high and low recruiting difficulty. Would such a merger average, increase, or lower subsequent recruiting difficulty for the new specialty? Similar questions arise in dealing with the impact of merging on subsequent attrition.

This paper is concerned with developing candidates for consolidation based on underlying principles and knowledge similarity of jobs. It is assumed that other considerations and analyses suggested here would be made following the similarity analyses made on underlying principles.

Electronic Principles Job Inventory

The Electronic Principles Job Inventory (EPI) and its development have been presented previously (O'Connor, Ruck & Driskill, 1978; Ruck, 1977) and therefore will be only briefly described in this paper. The EPI contains 1257 items covering the universe of electronic fundamentals as defined by Air Training Command fundamental courses (as of 1974) and instructors and supervisors of those courses. The 1257 items were written so that the job incumbent could indicate whether or not he or she uses each principle on the present job. Lead-in questions and routing instructions were provided to minimize the time required to complete the EPI booklet. For many sets of questions a "do not remember" question was included as an item after a list of detailed items was offered. This allowed the incumbent, for example, to indicate that he or she replaced capacitors on the present job but could not remember which type of capacitor was involved. Table 3 presents sample questions. It is important to note that the EPI was developed at the Occupational Measurement Center for the express purpose of course validation and was not originally intended to be a research tool. The Occupational

421

Measurement Center has collected EPI data from 59 specialites as of this writing.

Table 3
Sample EPI Questions

E1-1 Do you work with coupling devices in your present job? If no, go to item E2-1; if yes, continue.

Do you identify on schematic diagrams and relate to the actual circuitry the components associated with any of the following types of coupling?

E1-2 RC coupling
E1-3 Impedance coupling
E1-4 Transformer coupling
.
.
.

Do you work with any of the following types of coupling circuits?

E1-8 Directly coupled circuits
E1-9 Capacitive-resistive coupled circuits
E1-10 Capacitive-inductive coupled circuits
E1-11 Transformer coupled circuits
E1-12 Don't remember which type of coupling

Methods Used to Measure Commonality

The EPI was selected as the instrument to measure underlying principles/knowledge used within each specialty and, ultimately, as the input for commonality analysis. The EPI, for purposes of this analysis, is assumed to have included all of the relevant principles or knowledge required within the Air Force electronics community. Further, each item is assumed to have similar meanings across different specialties. Both assumptions are justified based on the development and validation procedures used in generating the instrument.

The criterion measure from the EPI that was selected was the percent of journeymen (5-skill level) personnel in each specialty answering "yes" to each item. Several difficulties arise in determining commonality once the criterion has been selected. First, no specialties have 100 percent overlap in principles used. Second, no measure of criticality (importance, difficulty, complexity, etc.) is currently available for inclusion in the analysis. Third, it is impractical to merge simple specialties with complex ones. Last, correlational measures could be quite misleading due to the possible high number of common zeroes. That is, correlations would be inflated due to the number of items that many specialties will have zero responses in common.

The statistical technique used in the analysis was Ward's hierarchical clustering technique (Ward, 1961). The sum of the absolute value of the differences in percent using over the 1257 items was input as the difference measure. The technique was employed so that common principles, the degree to which principles are used, and the size of each group being analyzed would be considered. Correlations among specialties that grouped in the cluster analysis were analyzed to provide additional interpretation of the overlap figures. Since no attempt was to be made to totally reorganize all electronic specialties, separate grouping analyses were performed for the 16 ground communications electronic specialties and the 4 wire and cable specialties.

Preliminary Results

Several specialties failed to group with other specialties in the pool. That is, they exhibited low overlap values with the most similar specialty or low correlations with the most similar specialty. Table 4 lists these specialties and pertinent EPI data. Two of the specialties, Telecommunication Systems/Equipment Maintenance (AFS 30652) and Telecommunications Systems Console Specialist/Attendant (AFS 30750) have very low utilization of electronic fundamentals, and would not appear to be good candidates for consolidation with any of the more complex specialties. The Television Equipment Repair Specialty (AFS 30455) appears to have low commonality with other specialties even though it has rather high utilization of fundamentals. Further detailed analysis is required for this specialty. Similarly, further analysis is required to determine why the Auto Tracking Radar Repair Specialty (AFS 30353) has low commonality in spite of moderate usage of principles.

Results of the grouping analyses are shown in Table 5. Since four specialties were omitted from this analysis, these results should be viewed as suggestive in nature. The groupings of specialties have been reviewed by Air Force managers and technicians in the ground communications electronics career field for their comments prior to performing additional analyses. It should be noted that the groupings displayed in Table 5 would be further analyzed using the additional considerations discussed earlier in this paper. Several of the groupings are congruent with recommendations that have already been made, and therefore validate prior judgments. Group E (Telephone Switching Equipment and Telephone Equipment Installation/Repair) has been formally proposed as a new specialty. Two of the three specialties in group B (Weather Equipment and Airborne Meteorological/Atmospheric Research) and in group D (Radio Relay Equipment and Ground Radio Communications Equipment) had also been formally proposed earlier. However, the third specialty in group B (Aircraft Control and Warning Radar) and the third specialty in group D (Space Communications Systems Equipment) had not been proposed as possibilities for merging within

423

Table 4
Specialties That Have Little Commonality with Other Communication Electronics Maintenance Specialties

AFSC	Title	Average Percent Used	Percent Used By Any	Percent Overlap with Most Similar Specialty	Correlation with Most Similar Specialty
30353	Auto Tracking Radar Repair	23	83	50	.68
30455	Television Equipment Repair	32	81	25	NA
30652	Telecommunications Systems/Equipment Maintenance	11	41	89	.38
30750	Telecommunications Systems Console Specialist/Attendant	4	21	89	.38

Table 5
Specialties That Have Potential for Consolidation

Group	AFSC	Title	Average Percent Used	Percent Used By Any	Percent Overlap Between Specialties	Correlation of Percent Members Performing
A	30554	Electronic Computer Systems Specialist	21	56		
	30651	Elec-Mech Comm & Crypto Equip Sys Spec	21	54	95	.94
B	30250	Weather Equipment Specialist	37	87		
	30352	AC&W Radar Specialist	37	89	93	.93
	30251	ABN Meteorological/Atmospheric Res Equip Repair	23	40	89	.77*
C	30451	Flight Facilities Equip Repair	33	84		
	30950B	Space Surveillance Radar Repair	29	74	94	.88
D	30450	Radio Relay Equipment Repair	23	74		
	30454	Ground Radio Comm Equip Repair	26	75	90	.85
	30456	Space Comm Sys Equip Opr/Spec	33	91	85	.83*
E	36251	Telephone Switching Equip, Electro Mechanical	7	27		
	36254	Telephone Equip Installation & Repair	6	24	91	.93
F	30351	Air Traffic Control Radar Repair	40	89		
	30950A	Missile Detection & Warning Radar	45	74	87	.83
G	36252	Electronic Switching Systems Repair	23	65		
	36253	Missile Control Comm Systems Specialist	13	40	67	.76

*Average correlation with each of the two preceding specialties those groups. Also, groups A, C, F, and G, although easily explained by knowledgeable electronics experts, had not been proposed, prior to this analysis, as new consolidated specialties.

The empirical groupings of specialties based on similarity of principles used as measured by the EPI have been supported, in some cases, by prior recommendations, and in other cases, by expert judgment. Although additional analyses are required prior to recommending implementation of the new consolidated specialties, the empirical procedures have provided an important service by reducing significantly the number of comparisons that should be made. In this study, for example, 190 comparisons of pairs of specialties would have to be made to examine all possible pairwise combinations, and 1140 combinations of 3 specialty groups must be studied in order to examine all groupings of that size. This analysis has narrowed the number of groups to be included in subsequent studies to only 7.

Plans for the Future

The results of this analysis indicate that the technique of grouping specialties using EPI data has considerable promise. Many of the groupings are logical and might have been expected. However, some of the groupings were not expected and require further detailed analysis. All of the potential consolidation groupings should be viewed as tentative and would be finalized only after additional analyses of occupational survey, manning, recruiting, and planning data have been performed. The contribution of the EPI data has suggested promising empirical groupings of specialties, something that was heretofore not possible. The EPI analysis has significantly reduced the scope of the comparison problem.

Several additional studies are planned in the near future. The analyses tentatively reported in this paper will be performed again once data for the complete array of specialties have been collected. Similar analyses will be performed for the total electronic community and within career areas within the electronic community.

Scales could be developed to further enhance the power of the EPI. Appropriate scales include measures of "complexity" or "difficulty" of the items. Once the scales have been developed, grouping could be performed on a measure of percent performing by (times) complexity. In addition, the new scale information would be quite useful to training specialists in course development.

Ultimately it might be useful to factor analyze the EPI so that a shorter non-redundant version could be used. Such analysis would be possible because there is considerable overlap among items in the inventory. This overlap is inherent, since the inventory covers knowledge, principle, task, and skill items of various degrees of specificity. The shorter version would be more efficient in terms of data collection and analysis, and would allow for "cleaner" grouping analyses.

The possible uses of EPI data have been enumerated elsewhere (O'Connor, Ruck & Driskill, 1978). Clearly, the utility of the instrument is, in large part, due to its universality. This paper has described one of many practical applications of the EPI data base. However, additional work is required, both in the development of analysis techniques, and in the refinement of the EPI.

References

O'Connor, T.J., Ruck, H.W., & Driskill, W.E. A universal model for evaluating basic electronic courses in terms of field utilization of training. Paper presented at the 86th Annual Convention of the American Psychological Association, Toronto, Canda, August 1978.

Ruck, H.W. (Ed). The development and application of the Electronic Principles Job Inventory. USAF Occupational Measurement Center Technical Note (7702). USAF Occupational Measurement Center, Lackland AFB, TX, October 1977.

Ward, J.H., Jr. Hierarchical Grouping to Maximize Payoff, WADD-TN-61-29. Personnel Laboratory, Lackland AFB, TX, March 1961.

SECTION 3

WOMEN IN THE ARMED SERVICES

Differential Field Assignment Patterns for Male and Female Soldiers

Laurel W. Oliver and Nehama Babin

US Army Research Institute for the Behavioral and Social Sciences

Paper presented at the meeting of the Military Testing Association, Oklahoma City, Oklahoma, November 3, 1978

Differential Field Assignment Patterns for Male and Female Soldiers[1]

INTRODUCTION

Background

Along with the other military services, the United States Army has been traditionally an almost totally male institution. Binkin and Bach (1977) have outlined the minimal role played by women in the military prior to World War II, their significant contribution during that war, and the consequences of the recent expanded role of women in the military. This expanded role, however, has not been accomplished without controversy. And the disputes regarding the role of women led to Army management's perception of the need for information on female enlisted personnel, particularly with respect to their performance.

Accordingly, since 1972, considerable attention has been given to assessing the effect of expanding the role of women in the Army. A number of different studies of women in the military have been conducted. Two major research efforts by the Army Research Institute for the Behavioral and Social Sciences (ARI) have concerned measuring the impact of female participation on performance. One of these investigations, called MAX WAC, involved a 72-hour field exercise and assessed the impact of varying levels of female content on unit performance (Army Research Institute, 1977). Later research, known as REF WAC, evaluated individual and group performance during an extended field training exercise (Johnson, Cory, Day, & Oliver, 1978).

Problem

During the REF WAC data collection, there were comments from the REF WAC participants concerning differential treatment of men and women. These subjective impressions were supported by the pretest and posttest questionnaire responses, which showed that sizeable proportions of respondents (officers, NCO's enlisted men, enlisted women) reported differential treatment of male and female soldiers by officers and NCO's. In general, about a third to more than a half of the respondents believed men and women were treated differently.

One mode of differential treatment may be the assignment of different jobs to men and women. Inspection of the REF WAC work availability data did, in fact, indicate that differential assignment patterns occurred during the field training exercise. Specifically, it was found that the mean number of regular work hours was greater for male enlisted personnel

---

[1] The authors wish to express their appreciation to Mr. Sidney Sachs for his invaluable contributions to the data analysis.

422

than for female enlisted personnel, while the mean number of special duty hours was greater for female enlisted personnel than for male enlisted personnel. Accordingly, it was decided to identify the differential patterns of regular and special duty assignments for men and women soldiers and to investigate the relationship of these patterns to possible causal variables.

One variable that might be related to differential assignment patterns is the mission of the unit. Since different types of units have different functions, assignment patterns could vary with the type of unit. And although the REF WAC data analysis showed that women had more special duty assignments than men, no analysis was made of the number of times special duty was assigned to each person, nor was the type of special duty broken out for men and women. Physical difficulty of the Military Occupational Specialty (MOS)[2] might be another factor affecting differential assignments, since the REF WAC results indicated that supervisors were strongly influenced by this variable in making hypothetical assignments to jobs. Finally, it was felt that a soldier's level of competence might affect the type of duty received. Performance ratings, then, might be related to assignment patterns.

Research Questions

The following research questions were investigated:

1. What are the relationships among gender, type of unit, and type of duty?

2. How is the frequency of special duty related to gender?

3. How is type of special duty task related to gender?

4. How is physical difficulty of DMOS related to assignment patterns?

5. How are daily performance ratings related to assignment patterns?

METHOD

The subjects, instruments, and procedures used in the REF WAC research project are described in detail in Johnson et al. (1978). Brief descriptions of the methodological aspects which pertain specifically to the research reported here are given below.

---

[2] In the Army, a Military Occupational Specialty (MOS) is a grouping of duty positions requiring similar qualifications and the performance of closely related duties. This job (e.g., clerk-typist, wheel vehicle mechanic) may be the soldier's primary one (Primary Military Occupational Specialty, or PMOS), his or her secondary one (Secondary Military Occupational Specialty, or SMOS), or the one which is his or her duty assignment (Duty Military Occupational Specialty, or DMOS). The DMOS may be the same as the soldier's PMOS or SMOS, or it may be an entirely different MOS.

433

Subjects

The population for this research included the enlisted personnel of 22 maintenance, medical, military police, signal, and supply and transportation units. These companies were among those participating in REFORGER 77, which involved an extended field training exercise conducted in West Germany in the autumn of 1977. The 22 units were selected because they contained women in sufficient numbers to provide a meaningful sample.

Members of the male and female cohorts of the REF WAC project constituted the sample for the research reported in this paper. All women in the selected companies were contained in the female cohort. A male from the same company was matched as closely as possible with each female on the basis of paygrade, length of service, MOS, age, and intelligence test score (see Johnson et al., 1978, p. II-10). These matched males constituted the male cohort.

Instrument

Data analyzed in this investigation were obtained from the Schedule 4 form, "Daily Record of Work Availability and Performance." This instrument, described in Johnson et al. (1978, pp. II-18 and II-21), was used to record supervisors' performance ratings and reports of work availability for each member of the male and female cohorts. The daily performance ratings were made on a seven-point Likert-type scale ranging from "performed all tasks in a superior manner" to "performed all tasks in an inferior manner." The work availability data consisted of records of assigned hours and lost time. Only the assigned hours for regular duty and for special duty were of interest to the research reported in this paper.

Data Collection

The Schedule 4 (Daily Record of Work Availability and Performance) data were collected by noncommissioned officers (NCO's) assigned to each company. Each day during the field training exercise, the NCO data collectors obtained performance ratings on each member of the male and female cohorts in their unit. The performance ratings were made by the individual's regular supervisor and/or the supervisor of the special duty to which the soldier was assigned. The number of hours on regular duty and on special duty was also recorded.

Design and Data Analysis

Variables. The independent variable which was the principal focus of this research was gender. That is, the primary comparisons were of male and female enlisted personnel. The other major independent variable of interest was type of unit (maintenance, medical, military police, signal, and supply and transportation). Both gender and type of unit were included in all analyses. (Due to small *n*'s, however, comparisons among units were not always meaningful.) For some analyses, the following independent variables were also included: duty assignment patterns (personnel with regular duty only, personnel with special duty and regular duty) and type of duty day (days with special duty, days with no special duty).

The dependent variables which were used in the analyses reported in this paper included the following:

1. Regular duty hours: mean number of hours per day spent on regular duty.

2. Special duty hours: mean number of hours per day spent on special duty.

3. Frequency of special duty: number of times a given individual was assigned to special duty.

4. Type of special duty: task assigned to soldier on special duty--vehicle maintenance (military police units only), guard duty, kitchen police, and "other" (see Johnson et al., 1978, pp. III-19 to III-21).

5. Difficulty of MOS: The physical difficulty of the Duty Military Occupational Specialty (DMOS) was evaluated according to the classification made in Johnson et al., 1978 (see pp. III-21 and III-22).

"1" = all tasks can be performed by a woman in a field environment.

"2" = most tasks can be performed by a woman in a field environment.

"3" = few tasks can be performed by a woman in a field environment.

If no DMOS was recorded, the subject's Primary Military Occupational Specialty (PMOS) was used. A few subjects had MOS which were not contained in the classification found in Johnson et al. (1978). In these cases, the MOS was evaluated by the second author in consultation with a senior military officer. For analysis purposes, categories "2" and "3" were combined. All MOS classified as "1" were defined as "easier MOS," and the remaining MOS were considered "harder MOS."

6. Performance ratings: mean ratings of performance for special duty or for regular duty.

Analyses. Several kinds of analyses were employed to explore the relationships among the variables described above. The principal analysis involved a repeated measures ANOVA which investigated the interrelationships of gender, type of unit, and type of duty. The two between-subjects factors were gender (male and female) and type of unit (maintenance, medical, military police, signal, and supply and transportation). The within-subjects factor was type of duty (regular and special).

In addition to the ANOVA, several chi square tests were conducted to examine the interrelationships among the variables. For the breakdown of special duty by frequency and type for male and female enlisted personnel, only the totals for all types of companies were used for the purpose of analysis. Although there appeared to be differences among the units, the n's were so small that more detailed analyses did not seem warranted.

A chi square test was also employed in assessing the relationship between special duty and MOS difficulty. The first test was of the number of males and females without special duty who were in harder and easier MOS. (Since inspection of the data revealed males and females with special duty were almost identically distributed in harder and easier MOS, no statistical test was run for this group.) The second chi square test involved the numbers of males and females (not broken down by whether or not they had had special duty) into harder and easier MOS.

No statistical tests were performed on the performance rating data. Means for the three types of duty days were so similar for men and women that further analysis was considered unnecessary.

426

RESULTS

Gender, Type of Unit, and Type of Duty

Table 1 contains means and standard deviations for regular and special duty hours for males and females by type of unit. The results show that for each type of unit males have more hours of regular duty than females. However, for special duty, with the exception of the military police units, females have more hours of special duty than the males. This pattern also appears in the totals for all males and all females.

Table 2 presents the results of the repeated measures analysis of variance. The results of the analysis revealed that all three main effects (for gender, type of unit, and type of duty) were significant beyond the .001 level. Two two-way interactions were also significant: type of duty by gender ($p < .001$) and type of duty by type of unit ($p < .001$). There was no significant interaction between gender and type of unit or for gender, type of unit, and type of duty.

Frequency of Special Duty

Table 1 shows differences between males and females for mean number of hours of regular duty and special duty for each type of unit. Table 3 presents a breakout of how many times individuals were assigned special duty in each type of unit. As can be seen in the table, most males and females who have special duty have it only one or two times. Of the five types of units, the military police and signal units stand out as the only ones in which the number of males who have one or two instances of special duty exceeds the number of females with one or two instances of special duty. It is also apparent from the table that of all soldiers who had special duty three or more times, there were considerably more women than men.

When a chi square test was performed on the total number of males and females who had special duty one or more times (see Table 4), there was a statistically difference between males and females for frequency of special duty ($x^2 = 6.97$, $p < .01$). This finding most probably was due to the fact that the number of females who had special duty three or more times was much greater than the number of males who had special duty three or more times.

Type of Special Duty

When investigating patterns of special duty assignments, it is of interest to ascertain what kind of tasks are assigned for special duty and if these tasks differ for men and women. It can be seen in Table 5 that most instances of special duty were either guard duty or kitchen police. The pattern of more special duty for women that was previously noted can be observed here also. In three types of units (medical, maintenance, and supply and transportation), women have more instances of guard duty and kitchen duty than men. However, a chi square test (see Table 6) revealed no statistically significant difference between the number of instances of guard duty and kitchen police duty between males and females ($x^2 = .002, p > .05$).

Difficulty of MOS

The totals in Table 7 indicate that a larger proportion of both males and females with special duty were in easier MOS (72% and 71% respectively), than were males and females with no special duty (38% and 51% respectively). There was virtually no difference between the proportions of men and women with special duty in either easier or harder MOS. The pattern differed, however, for personnel with no special duty. For these individuals, males tended to be concentrated in the harder MOS (62% males vs. 49% females), and a chi square test ($x^2 = 5\ 05$, df = 1) (see Table 8) showed this difference to be significant at $\alpha = .05$.

In spite of the matching procedure followed in selecting the male cohort, Table 9 demonstrates that there was a significant difference in MOS difficulty between males and females, with males tending to be in harder MOS and females in easier MOS. This difference is significant at $\alpha = .05$ ($x^2 = 5.0$, df = 1).

Performance

Table 10 presents mean performance scores for males and females in each type of unit for three different types of duty days: (1) days of special duty for personnel with special duty, (2) days of regular duty for personnel with special duty, and (3) days of regular duty for personnel with regular duty only. As can be observed in the table, the data do not suggest any relationship between performance ratings and type of duty day. Therefore no statistical test was performed.

DISCUSSION

Limitations

The findings of this research must be considered with an awareness of the limitations of the investigation. First, the data collection was constrained by the general requirement of noninterference in the normal activities of the subjects. An additional problem was the lack of time for the training of data collectors with consequent adverse effects on reliability. In retrospect, it appeared that the definitions of terms and categories were not always clear to the data collector and the supervisors from whom they collected the data. The result of this lack of consistent definition was that different units (and different data collectors) recorded time in different ways. Tasks assigned as regular duty in one unit, for example, might be considered special duty in another. Shifts in one type of unit might be eight hours in length, while in another unit shifts were considered 24 hours long because the "on call" time was included. In addition, there were the usual individual differences among data collectors contributing additional variance.

Because in many cases the data were based on small _n_'s, statistical comparisons could be made only for totals. Yet it is possible that comparisons across different types of units may not always be meaningful due to differences in unit mission or to differences in classification of the performance variables, etc. Hence, generalizations from these results should be made cautiously. The research does, however, demonstrate interpretable trends and suggest lines of investigation for future research.

Conclusions

Gender, type of unit, and type of duty. Of the three significant main effects found in the repeated measures ANOVA, the one for gender was of greatest interest. It can be concluded that, on the average, men worked significantly more hours than women during the field training exercise. The main effect for type of unit showed that the total amount of time worked by enlisted personnel varied significantly as a function of the type of unit. These differences in total time among units may be due, at least in part, to differences in the units' missions. As noted above, however, some of the difference may have been due to variations in interpreting terms. For example, some supervisors considered time "on call" as part of the regular shift and other supervisors did not. That the daily average for regular duty hours would differ significantly from the special duty average was obvious from visual inspection of the data and not of interest in and of itself.

The pattern of differential assignment was revealed by the significant interaction between gender and type of duty. It was clear that females had significantly more special duty and males had significantly more regular duty during the field exercise. This pattern was noted for [illegible] units for both types of duty except for military police units in which women had less special duty. The interaction of type of unit and type of duty suggested that units varied in terms of the relative amount of regular and special duty assigned. Again, this finding may have been affected by differences in recording data. However, no interaction occurred between type of unit and gender. Thus, it can be concluded that the amount of total time assigned to men and woman did not vary as a function of type of unit. The three-way interaction was not significant, either, indicating no reliable differences beyond those occurring in the two-way interactions.

Frequency of special duty. Because the n's are small, no conclusions can be reached concerning the frequencies of special duty for different types of units (Table 3). Overall, however, it is clear that women not only had significantly higher daily averages for special duty than men (as demonstrated by the ANOVA results in Table 2) but that they tended to be assigned to it more often than men (Table 4).

Type of special duty. When special duty was assigned, the most frequent kinds were guard duty or kitchen police duty. A chi square test showed that the numbers of men and women assigned to these jobs did not differ significantly. Therefore, it can be concluded that no bias seemed to be operating in the type of jobs assigned to the special duty soldiers. That is, men did not tend to draw guard duty, and women were not necessarily put on kitchen police duty.

Difficulty of MOS. Difficulty of MOS appeared to have no relationship to gender for personnel who had been assigned special duty since virtually identical proportions of males and females fell into the "harder" and "easier" MOS classifications. For personnel with no special duty, however, the pattern differed significantly, with more men concentrated in the harder MOS category. As is apparent from Table 7, almost three-fourths of the people with special duty had easier MOS. Females with no special duty tended to be evenly balanced between harder and easier MOS, while 62% of men with no special duty were in harder MOS. This difference (see Table 8) proved to be significant ($x^2 = 5.05$, df = 1, $p < .05$). The pattern shown in Table 7 suggests that supervisors tended to assign people from easier MOS to special duty. It may be that people in easier MOS are more interchangeable and that those in harder MOS are difficult to replace due to the physical demands of the job. It would make sense, then, for a supervisor to assign the more easily replaced soldier to special duty. One interpretation of these findings is that women may get special duty more often not because they are women but because they are in easier MOS. (Women may, of course, be concentrated in easier MOS to begin with because of gender bias.)

Performance. It was felt that performance might be related to special duty with lower performers being selected for this type of duty because, like soldiers in easier MOS, they would be more easily replaced or compensated for. The results showed, however, that there was no relationship between performance ratings and type of duty for either those who had had only regular duty or those who had been assigned to special duty one or more times.

Implications

Perhaps the principal contribution of this research is to illustrate that investigations of male-female differences should not examine the gender variable in isolation. The obvious conclusion when significant differences are found between male and female groups (especially in the Army) is that bias is operating. Yet such is not necessarily the case. Here, for example, women receive significantly more special duty than men. But of those persons assigned special duty, there was no difference between males and females in type of duty assigned. The traditional male job (guard duty) was no more likely to be assigned to males than to females, and the same pattern held for the traditional female job (kitchen police duty). Analyses related to MOS difficulty suggested that it may have been this variable, rather than gender, which was important in the selection of people for special duty. Special duty personnel tended to come from the easier MOS. And although there was an attempt to match males and females on the MOS variable, women may have been overrepresented on special duty because there was a significant difference in the way men and women were assigned to MOS. In addition to being an example of the need to investigate overall gender differences, this research has additional implications as described below.

Future research. The findings described in this paper have implications for future research on assignment patterns. If similar data are collected again, every effort should be made to obtain more precise definitions of the categories to be used. "Regular duty," for example, should draw from the same set of behaviors for every subject. Data collectors should receive sufficient training in defining terms and behaviors and in recording data, and they should be able to transmit this knowledge to the supervisors. It would be important to know who was doing what, when and why they were doing it, and for how long and under what circumstances. The leadership aspect of assignments should be delineated in detail--who makes what assignments for whom, and why these people were selected. The Schedule 4 performance ratings used in this research were found to have some validity in the REF WAC research since two other REF WAC observational performance measures yielded similar findings concerning male-female performance (Johnson et al., 1978). But reliability data concerning the duty categories are lacking, and such data should be provided for future research efforts.

In some of the data analyses described in this paper, comparisons could not be made within each type of unit because of the small number of instances recorded. Since the missions of different types of units vary, it may not be meaningful to combine data across units. Hence, future research should endeavor to collect enough data within each unit type to permit intra-type comparisons.

411

Male-female assignments. This research has relevance for those concerned with patterns of male and female assignments. The findings suggested that women got more special duty during the extended field training exercise not because they were women but because they were in easier MOS. The results also showed that men worked significantly more hours than women. If the proportion of women in the Army (or in certain Army units) rises markedly, men may be increasingly concentrated in the harder MOS. If, at the same time, men must work more hours (at least, during field exercises), morale could be adversely affected and the ability to accomplish the unit mission may decrease. Differential assignment patterns, then, are potentially detrimental and may impair organizational effectiveness.

SUMMARY

This paper has examined assignment patterns of male and female enlisted personnel during a field training exercise. Although female enlisted personnel averaged a greater amount of special duty per day than did their male counterparts, men averaged more regular duty and more total duty per day than did women. Women were found to be more frequently assigned to special duty than were men. However, of those people assigned to special duty, there was no discrimination in terms of the type of duty assigned to men and women. Women were as likely to have guard duty as men, and men were as likely to have kitchen police duty as women. It was also found that significantly more women than men had easier MOS. Of those personnel assigned to special duty, almost three-fourths had easier MOS; the pattern was identical for men and women. Of personnel never assigned to special duty, less than one-half were in easier MOS; a greater concentration of men than women were in harder MOS. The findings suggested that supervisors tended to select special duty people from those with easier MOS, perhaps because persons in easier MOS were more interchangeable and their absence was more easily compensated for. Since the matching of the male and female cohorts on MOS was imperfect, it could not be determined whether the differential assignment pattern was due to the overconcentration of women in easier MOS or to gender bias. Unlike MOS difficulty, performance ratings proved to be unrelated to special duty. The conjecture that lower performers would tend to be assigned to special duty was not confirmed.

Thus, the primary contribution of this research was seen as an illustration that investigations of male-female differences cannot consider the gender variable in isolation. The implications of the research were also discussed in terms of methodological considerations for future research and the effect of differential assignment patterns on organizational effectiveness.

References

Army Research Institute. Women content in units force development test (MAX WAC). (ARI Special Report S-7). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 1977.

Binkin, M., & Bach, S.J. Women and the military. Washington, D.C.: Brookings Institution, 1977.

Johnson, C.D., Cory, B.H., Day, R.D., & Oliver, L.W. Women content in the Army - REFORGER 77. (ARI Special Report S-7). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 1978.

Table 1

Means and Standard Deviations for Regular and Special Duty Hours for Enlisted Personnel

Type of Unit	N	Duty Hours				
		Regular Duty		Special Duty		Total Duty
		M	SD	M	SD	
Maintenance						
Male	57	12.68	2.91	.55	1.08	13.23
Female	56	12.59	3.36	.86	3.79	13.45
Medical						
Male	21	9.73	3.82	.23	.36	9.96
Female	19	7.41	2.84	.71	.07	7.81
Military Police						
Male	39	11.04	1.62	.29	.56	11.33
Female	42	9.80	1.95	.16	.34	9.96
Signal						
Male	31	12.42	2.24	.14	.31	12.73
Female	31	10.92	1.98	.21	.59	11.13
Supply & Transportation						
Male	47	11.52	2.46	.36	.89	11.88
Female	54	10.11	3.08	1.45	3.02	11.56
Totals						
Male	195	11.71		.35		12.06
Female	202	10.60		.76		11.36
Entire Sample	397	11.15		.56		11.71

Table 2

Summary of Analysis of Variance for Gender, Type of Unit, and Type of Duty

Source of Variance	Degrees of Freedom	Mean Square	*F* Ratio
Between subjects			
Gender (A)	1	38.57	12.65***
Unit (B)	4	84.65	27.76***
	4	7.22	2.37 n.s.
Within subjects			
Duty (C)	1	18423.17	2886.36***
CA	1	121.36	19.01***
CB	4	69.92	10.96***
CAB	4	9.87	1.55 n.s.

*** $p < .001$

Table 3

Frequency of Special Duty for Each Male and Female Soldier by Type of Unit

Type of Unit	Frequency			
	0 times	1-2 times	3 or more times	Total
Maintenance				
Male	41	15	1	57
Female	32	20	4	56
Medical				
Males	14	7	0	21
Females	10	6	3	19
Military Police				
Male	28	11	0	39
Female	32	9	1	42
Signal				
Male	24	7	0	31
Female	25	5	1	31
Supply & Transportation				
Male	40	5	2	47
Female	35	11	8	54
Total				
Male	147	45	3	195
Female	134	51	17	202

Table 4

Frequency of Special Duty for Males and Females

Group	Frequency of Special Duty		
	1-2	3 or more	Totals
Males	45(39.72)[a]	3(8.27)	48
Females	51(56.28)	17(11.72)	68
Totals	96	20	116

$x^2 = 6.97 \; \underline{p} < .01$

---

[a] Expected frequencies in parentheses

Table 5

Frequency of Type of Special Duty Tasks Assigned to Enlisted Personnel

Type of Unit	Type of Task Vehicle Maintenance	Guard Duty	Kitchen Police	Other[a]	Total
Maintenance					
Males		4	19	1	24
Females		5	31	1	37
Medical					
Males		6		4	10
Females		13	3	2	18
Military Police					
Males	6		6	0	12
Females	5		4	4	13
Signal					
Males		6	1	1	8
Females		4	3	0	7
Supply & Transportation					
Males		20	5	2	27
Females		36	12	3	51
Total					
Males	6	36	31	8	81
Females	5	58	53	10	126

---

[a] This category includes unexplained tasks and tasks described generally as "details."

Table 6

Number of Instances of Guard Duty and Kitchen Police Duty Assigned to Enlisted Personnel

Enlisted Personnel	Type of Duty Guard Duty	 Kitchen Police	Totals
Males	37 (36.3)[a]	32 (32.2)	69
Females	59 (59.2)	52 (51.8)	111
Totals	96	84	180

$x^2 = .002$, $p > .05$

[a] Expected frequencies in parentheses

Table 7

Numbers and Proportions of Males and Females in Easier and Harder MOS by Type of Unit

Type of Unit[a]	Personnel with Special Duty				Personnel with No Special Duty			
	Easier MOS		Harder MOS		Easier MOS		Harder MOS	
	n	%	n	%	n	%	n	%
Maintenance								
Males	12	75	4	25	22	48	24	52
Females	15	65	8	35	28	64	16	36
Medical								
Males	7	88	1	13	11	85	2	15
Females	9	100	0	--	7	64	4	37
Signal								
Males	4	58	3	43	5	12	38	88
Females	3	50	3	50	12	28	31	72
Supply & Transportation								
Male	8	67	4	33	15	38	24	62
Female	13	72	5	28	20	61	13	39
Totals								
Male	31	72	12	28	53	38	88	62
Female	40	71	16	29	67	51	64	49

[a] MP MOS data not available

Table 8

Percent of Easier and Harder MOS for Males and Females with No Special Duty

Enlisted Personnel	Type of MOS		
	Easier	Harder	Totals
Males	53(62.2)[a]	88(78.8)	141
Females	67(57.8)	64(73.2)	131
Totals	120	152	272

$x^2 = 5.05$ $p < .05$

---

[a] Expected frequencies in parentheses

Table 9

Total Number of Males and Females with Easier and Harder MOS

Enlisted Personnel	MOS		
	Easier	Harder	Totals
Males	84 (94.7)[a]	100 (89.3)	184
Females	107 (96.3)	80 (90.7)	187
Totals	191	180	371

$$x^2 = 5.0 \ \underline{p} < .05$$

[a] Expected frequencies in parentheses

Table 10

Mean Performance Scores of Enlisted Personnel for Three Types of Duty Days

	Personnel with special duty						Personnel with regular duty only		
Type of Unit	Days with special duty			Days with no special duty					
	n	M	SD	n	M	SD	n	M	SD
Maintenance									
Males	16	6.4	1.00	16	6.1	.80	41	6.0	1.06
Females	24	5.8	1.57	23	5.8	.92	32	6.0	.70
Medical									
Males	7	5.6	.73	7	5.9	.52	14	5.4	1.02
Females	9	6.1	1.05	9	6.0	.78	10	5.6	1.23
Military Police									
Males	11	5.5	1.77	11	5.8	.93	28	6.0	.76
Females	10	6.0	.48	10	5.9	.73	32	5.9	.73
Signal									
Males	7	6.3	.45	7	5.7	.90	24	5.9	.65
Females	6	6.6	.56	6	5.4	.66	25	5.7	.77
Supply & Transportation									
Males	7	5.1	.76	7	5.3	1.22	40	5.8	.69
Females	19	5.9	1.09	19	6.2	1.10	35	5.9	.82
Totals									
Males	48	5.9	1.23	48	5.8	.92	147	5.9	.86
Females	64	6.0	1.21	67	5.9	.95	134	5.9	.81

THE PREMATURE ATTRITION OF NAVY FEMALE ENLISTEES*

Gerry L. Wilcove
Patricia J. Thomas
Constance Blankenship

Navy Personnel Research and Development Center
San Diego, California 92152

*Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, 30 October-3 November 1978.

THE PREMATURE ATTRITION OF NAVY FEMALE ENLISTEES[1]

Background

Gradually, women are obtaining greater opportunities in the Navy defense establishment, as evidenced by progress in four areas: (1) it is easier for women to gain entrance into the Navy and Navy programs than it has been previously, (2) women are participating in a greater range of work activities, (3) a greater number of training opportunities are open to them, and (4) there are signs that they are being accepted by Navy management as "one of the Navy's own."

Four historical developments indicate that it is becoming progressively easier for women to gain entrance to the Navy and its programs. In 1967, the two percent ceiling on women in each service was abolished by Congress. In 1972, Admiral E. Zumwalt, the then Chief of Naval Operations, made women eligible for NROTC scholarships, thereby giving them access to another officer training program. In 1974, the age regulations governing selection procedures throughout the military were standardized so that women were no longer required to be older than men. In 1976, entrance of women into the military academies was cleared by an amendment to the Defense Appropriation Bill.

One development, in particular, illustrates the expansion of women's work roles in the Navy, i.e., issuance of "Z-Gram 116" by Admiral Zumwalt in 1972. This directive specified that women were to become eligible for (1) all enlisted ratings, (2) shore command positions, (3) the Chaplain and Civil Engineer Corps, and (4) flag rank (i.e., admiral status) within managerial and technical specialties.

One indication that women are receiving greater training opportunities in the Navy was the issuance in 1976 of a directive requiring women to take apprenticeship training if they were not eligible for "A" School. This training provides a basic shipboard orientation and is a prerequisite for acquiring an apprenticeship position.

In addition to the expansion of women's work roles and training opportunities, there are signs that women are gradually being accepted by Navy management as "one of the Navy's own." For example, their uniform has been redesigned to make it more compatible with their new work roles. Also, a "summer whites" uniform, previously available only to men, is currently being tested for women.

In summary, it can be stated that women increasingly are becoming a more integral part of the Navy community and are shouldering a greater share of the defense burden. Moreover, this trend is expected to continue--the Chief of Naval Personnel (Note 1) has recommended that the Navy double its percentage of women by 1983.

---

[1] The opinions and assertions contained herein are those of the senior author and are not to be construed as official or reflecting the views of the Navy Department.

Problem

Two problems gave rise to the current study. First, although the attrition rate of Navy female first enlistees has been declining since 1973, it is still considered to be too high--approximately 28 percent (see Thomas, Note 2). Secondly, it is expected that the attrition rate will increase when and if women are no longer required to have a high school diploma for acceptance into the Navy, i.e., when and if selection requirements for women become the same as those for men. Research with men (Plag & Goffman, 1966; Lockman & Gordon, 1977; Sands, 1977) has consistently demonstrated that educational level is the most valid predictor of premature attrition.

Purpose

The current study was an exploratory one, designed to lay the groundwork for (1) an instrument which could be used in the relatively near future for screening female applicants (Goal 1) and, to a lesser extent, (2) a screening instrument which could be used when and if female applicants are no longer required to have a high school degree (Goal 2). In order to reach these goals, the study investigated the relationship between the premature attrition of Navy female first enlistees and preenlistment variables, such as personal history, demographic traits, and attitudes. It was believed that these variables would be especially useful for reaching Goal 1. That is, it was believed that these variables would be less attenuated for women, and thus more predictive, than, for example, most of the variables currently used to select males (Lockman & Gordon, 1977). The selection procedure for males utilizes age, mental level, educational level, and number of dependents for which an applicant is financially responsible. However, the last three variables are attenuated for women; mental level and educational level because of the high school degree requirement and the last because the male typically assumes responsibility for a family's financial obligations. It should be noted that educational level and mental level should become more useful as predictor variables when and if the high school degree requirement is abolished for women. It was believed, however, that additional variables will be needed in order to effectively predict attrition. The present study (Goal 2) laid the groundwork for locating them.

Approach

Instruments. A questionnaire approach was utilized in the study. Two questionnaires termed Quest 1 and 2 were constructed, composed of "common" and "unique" items. Common items were those which were identical on both questionnaires. Unique items were those which were found on one questionnaire but not the other. The unique items on one questionnaire were, at times, parallel in form to those on the other questionnaire. At other times, unique items on one questionnaire measured totally different aspects of a general construct, such as mental health, than did items on the other questionnaire.

Each questionnaire was composed of 120 items which were conceptually grouped into eight areas, as shown in Table 1. Copies of Quest 1 and 2 are available from the senior author.

Table 1

Content Areas and Number of Items: Quest 1 and 2

Content Area	Number of Items
Personal History/Demography	26
Female Role Ideology	6
Mental Health	24
Motivation to Fail	6
Realistic Expectations of Navy	4
Enlistment Motivation	21
Similarity to Previous Successful Recruits	6
Occupational Needs	27
TOTAL	120

Items relating to personal history and demographic traits were conceived on a logical basis, i.e., they were perceived as being related to attrition. Item topics included number of males in the household when growing up, previous emotional reactions to time spent away from home, and type of discipline received during teenage years.

Items were included on female-role ideology, using the concepts of "traditional" and "contemporary" ideologies advanced by Lipman-Blumen (1972). Approximately 80 percent of the enlisted women in the Navy are assigned to traditional jobs. It was thus hypothesized that the less traditional a woman was in her role orientation, the more likely she was to attrite.

It was believed in the current study that poor mental health was related to attrition ((see, for example, Craighill (1947) and Schuckit and Gunderson (1971)). Mental health items developed by Friedman (1956) were used in Quest 1 and 2.

458

Horner (1969) found that the motive to fail characterizes the personality of many women. It was believed in the current study that this motive would lead ultimately to attrition. A story involving a hypothetical, successful, woman Recruit Chief Petty Officer was included in the questionnaires, as a projective device, to measure the motive to fail.

Porter and Steers' review article (1973) concluded that one cause of turnover may be the unrealistic expectations of individuals upon entering an organization. Items were thus constructed which asked the respondent whether she had any relatives or friends in the military and whether she had discussed their experiences with them.

Enlistment motivation was also tapped by the questionnaires. To develop items, 50 women were interviewed who had recently been assigned to their first duty station. Generally speaking, an "empirical" approach was utilized, i.e., no hypotheses were advanced relating enlistment motivation to attrition.

Items were designed to assess the similarity of the respondent to previous successful recruits. A woman who had previously been a recruit company commander was interviewed and "success" traits identified, such as (1) a tendency toward conformity, (2) a commitment to the Navy as a career, and (3) a deliberative rather than an impulsive decision-making style.

It was believed in the current study that a general set of occupational needs may exist that are optimally compatible with Navy life. Individuals possessing these needs would be more likely to experience job satisfaction and perhaps less likely to attrite. Items were adapted for use from Hall's Occupational Orientation Inventory (1971).

Sample and Data Collection. One of the two questionnaires was administered to each female recruit after she had been assigned to her company, but before her actual training had begun.[2] Twenty companies participated, with a total N of 977. Questionnaires were administered in May, June, and July of 1975 and a deadline of December 1976 established for determining whether a woman was an attritee or a survivor.[3]

Data Analysis. Analyses were conducted for individual items (referred to below as the "first" and "second" analyses) and scales (referred to as "third" and "fourth" analyses.) In the *first* analysis, chi-square and

---

[2] Since the study was an exploratory one, the "restriction of range" problem inherent in using a recruit sample was not deemed critical.

[3] The use of one deadline for all subjects, instead of, for example, using an 18-month lag period for each person, was not viewed as serious because of the study's exploratory nature.

"strength of association" (SA) statistics were computed between each item and attrition. The Cramer *V* was utilized as a measure of SA for nominally-scaled items, while tau *b* and *c* were used for ordinally-scaled items. Sample *N*'s for the unique Quest 1, unique Quest 2, and common items were 485 (105 attritees and 380 survivors) (*Sample A*), 492 (99 attritees and 393 survivors) (*Sample B*), and 977 (204 attritees and 773 survivors) (*Sample C*), respectively. As reported later, more unique Quest 1 items emerged as significant in the chi-square analysis than did for the other types of items. Therefore, in the *second* analysis, Sample A (1) was divided randomly in two, (2) ordinally-scaled items were identified in the first subsample which evidenced a tau of .10 or greater, and (3) a regression analysis was conducted with the second subsample utilizing (a) the identified items as predictors and (b) the attritee--survivor status of the woman as the criterion. A shrunken *R* was then computed.

In the *third* analysis, the empirical keying approach of Campbell (1971) was utilized to identify a set of discriminating response options or "scale", i.e., a set of response options, each of which had been selected by at least 10 percent more attritees than survivors, or vice versa. This approach was utilized separately for Samples A, B, and C. The *fourth* analysis also utilized Campbell's approach. Samples A, B, and C were each randomly divided into a validation and cross-validation sample. A unit weighting scoring system was devised based on validation sample data. That is, a plus or a minus 1 was assigned to a discriminating response option, as appropriate, and a zero to non-discriminating options. This system was then utilized with the cross-validation sample to produce a scale score for each woman. Scores for the entire cross-validation sample were then correlated with attrition.

RESULTS

The Relationship Between Items and Attrition

Table 2 presents the results for all the unique Quest 1 items which were significantly related to attrition. These items are broken down into those which were ordinally-scaled and those which were nominally-scaled. For ordinally-scaled items, the exact nature of their relationship with attrition is specified--for example, the table indicates (see Item 41) that a woman is more likely to attrite if she values individuality as opposed to conformity. For the nominally-scaled items, only the general content of the item is supplied.

Twenty-one of the 57 items unique to Quest 1 demonstrated a statistically significant relationship with attrition in the chi-square analysis ($p < .10$).[4] However, the absolute strength of these relationships was weak, obtained statistics varying from .008 (item 92, tau *c*) to .215 (item 52, tau *b*). Of the eight areas advanced at the start of the study as possible indicators of attrition, only two--mental health and occupational needs--produced a sizeable number of significant items in the chi-square analysis. The relationship between the mental health items and attrition was as hypothesized, i.e., the more the woman perceived herself as nervous, headache-prone, etc., the more likely she was to attrite. Although no hypotheses were advanced for the occupational need items, a discernible need profile emerged for the attritee.

Table 3 presents the results for those items unique to Quest 2 which were significantly related to attrition, while Table 4 presents the results for significant common items. Eight of the 57 unique Quest 2 items and nine of the 63 common items were significantly related to attrition, although, once again, the actual strength of these relationships was weak. Most of the significant unique Quest 2 items represented the mental health area, while the significant common items represented the personal history and enlistment motivation areas.

Even though strength of association statistics were generally low for the significant items, the possibility existed that combining the items in a multivariate fashion would increase their predictive value. An exploratory analysis was conducted, therefore, composed of the following steps: (1) Individuals in Sample A (i.e., individuals completing Quest 1) were randomly assigned in a 50-50 fashion to one of two subsamples, (2) ordinally-scaled items were identified in subsample 1 which evidenced a tau statistic $\geq .10$, and (3) these items were then utilized as predictors with subsample 2 in a multiple regression analysis in which the survivor-attritee status of the woman served as the criterion. Eighteen items were subsequently identified for subsample 1. Results are available from

---

[4] It was judged that this significance level was appropriate for an exploratory study.

Table 2

Unique Quest 1 Items Which Were Significantly Related to Attrition

Ordinally-Scaled Items

Item Number	Category[a]	X^2(df)	p	Strength of Association[b]	Nature of Relationship: Woman More Likely to Attrite If She
41	SR	11.109(3)	.011	-.067	Values individuality
46	MH	13.903(1)	.001	.174	Reacts poorly to pressure
47	MH	5.261(1)	.022	.109	Is nervous person
52	MH	21.087(1)	.001	.215	Experiences many headaches
56	MH	3.637(1)	.057	.097	Had bossy teachers
60	MH	5.140(1)	.023	.109	Is subject to depression
62	MH	19.687(1)	.001	.208	Trembles with anxiety
63	MH	5.366(1)	.021	-.114	Had poor childhood health
64	MH	7.579(1)	.006	.131	Has difficulty sleeping
66	MH	12.809(1)	.001	.168	Worries a lot
71	ON	12.908(4)	.012	-.105	Desires autonomy
73	ON	8.981(4)	.062	.109	Doesn't value novelty
76	ON	16.488(4)	.002	.106	Doesn't value work teams
77	ON	9.185(4)	.057	.064	Doesn't value caring supervisor
78	ON	8.698(4)	.069	.083	Doesn't value job respect
79	ON	14.376(4)	.006	.083	Doesn't value orderly procedures
89	ON	8.355(4)	.079	.031	Doesn't value set plans
92	ON	8.083(4)	.089	.008	Doesn't value interpersonally-oriented jobs
94	ON	13.060(4)	.011	-.089	Enjoys working outdoors

Nominally-Scaled Items

Item Number	Category[a]	X^2(df)	p	Strength of Association[b]	Item Content
34	RE	19.630(4)	.001	.202	Group activities with males
35	FR	10.492(4)	.033	.150	Family religion

Note. N = 485.

[a] Category abbreviations: RE = items on realistic expectations about Navy, FP = items on female role ideology, SR = items on similarity to previous successful recruits, MH = mental health items, ON = occupational need items.

[b] For ordinally-scaled items, tau b was computed for the 2 x 2 situation, i.e., when the df were 1, and tau c was computed in all other situations, i.e., when the df were greater than 1. For nominally-scaled items, a Cramer V was computed.

Table 3

Unique Quest 2 Items Which Were Significantly Related to Attrition

Item Number	Category[a]	X^2(df)	p	Strength of Association[b]	Nature of Relationship: Woman More Likely to Attrite If She
34	MH	10.694(4)	.030	-.054	Had a neglectful mother
47	MH	5.619(1)	.018	.112	Is nervous
49	MH	6.445(1)	.011	.121	Had severe childhood punishment
52	MH	5.252(1)	.022	.110	Faints a lot
54	MH	6.574(1)	.010	.122	Believes she has bad luck
60	MH	9.112(2)	.011	.099	Is chronically tired
66	MH	6.753(1)	.009	.123	Becomes upset when yelled at
93	ON	10.558(4)	.032	-.002	Wants to travel

Note. N = 492. All items were ordinally scaled.

[a] Category abbreviations: MH = mental health items, ON = occupational items.

[b] tau b was computed for the 2 x 2 situation, i.e., when the df were 1, and tau c was computed for all other situations, i.e., when the df were greater than 1.

Table 4

Common Items Which Were Significantly Related to Attrition

Ordinally-Scaled Items

Item Number[a]	Category[b]	X^2(df)	*p*	Strength of Association[c]	Nature of Relationship: Woman More Likely to Attrite If She
19	PH	9.688(4)	.046	.092	Dates infrequently
21	PH	15.578(4)	.004	.088	Plans to marry/remarry
111	EM	9.759(4)	.048	-.055	Doesn't want to travel/meet people[d]
113	EM	8.003(4)	.092	-.028	Doesn't want further education
115	EM	14.238(4)	.007	.096	Has relatives/friends in service
117	EM	18.138(4)	.001	.110	Wants to help family financially

Nominally-Scaled Items

Item Number[a]	Category[b]	X^2(df)	*p*	Strength of Association[c]	Item Content
14	PH	12.240(4)	.016	.124	Childhood clubs
17	PH	15.038(4)	.005	.124	Types of male friendships
20	PH	16.727(4)	.002	.131	Marital history

Note. *N* = 977.

[a] Item numbers were the same for both questionnaires.

[b] Category abbreviations: PH = personal history items, EM = enlistment motivation.

[c] All values for the ordinally-scaled items are tau *c* statistics. Cramer *V*'s are entered for the nominally-scaled items.

[d] This item varied somewhat in content from Item 93 in Table 3, perhaps accounting for the different results.

the senior author. A multiple R of .388 was obtained with subsample 2 and a shrunken R of .295 (Guilford & Fruchter, 1978, p. 377).[5]

The Relationship Between Scales and Attrition

As described previously, response options were identified for Samples A, B, and C which discriminate between attritees and survivors. In any future studies, these options are likely to be the most stable since the entire sample was utilized in each case. Information on these options is available from the senior author.

To claim that these options are the most stable is not sufficient, however: They may not be stable enough to base a screening instrument on them. Therefore, an attempt was made to obtain some quantitative information. As described previously, Samples A, B, and C were each divided into a validation (V) and cross-validation (CV) sample. A set of discriminating response options, or scale, was identified for each V sample, the three scales respectively termed (1) the Unique Quest 1 Scale, (2) the Unique Quest 2 Scale, and (3) the Common Scale. Options identified in the V sample were then unit weighted in the CV sample and a scale score computed for each woman. Correlating such scores with attrition yielded cross-validation coefficients of .247, .149, and .069 for the above-mentioned Unique Quest 1, Unique Quest 2, and Common Scales, respectively.

---

[5] Although restricted samples were utilized, one would generally expect the reported correlations to be underestimates.

405

CONCLUSIONS AND RECOMMENDATIONS

As reported, a shrunken $\underline{R}$ of .30 was obtained between attrition and a set of items measuring preenlistment variables. Moreover, a cross-validation correlation of .25 was obtained between attrition and one of the scales created through empirical keying. Both results suggest that a moderately effective instrument can be constructed for screening current female applicants, although the regression approach seems more promising.

If one examines Tables 2, 3, and 4, one sees that there are 25 items with "strength of association" values $\geq$ .10. At a time when female applicants are required to have a high school education, these items represent the most logical choices for a selection instrument. It is recommended, however, that researchers first evaluate these items from a strict legal and practical standpoint. Some items--for example, those on occupational needs--pose no obvious problem. However, an item on family religion could not ethically or legally be used. There may be some question about using the mental health items, because they may represent an invasion of privacy. Also, it may be possible to "fake" one's responses to the mental health items. That is, if applicants are informed, as required by the Privacy Act, that these items are being used to screen them, they may falsify their answers.

Items which survive this evaluation should then be administered to female applicants at the Armed Forces Entrance and Examination Centers, along with the Armed Services Vocational Aptitude Battery (ASVAB). It is believed by the authors that the Navy is committed to using the ASVAB as its primary screening device for the foreseeable future. The goal in the proposed study, therefore, becomes one of determining whether "preenlistment" questionnaire items significantly improve attrition prediction over and above the ASVAB. (It was impossible in the current study to include the ASVAB as a variable, since the Basic Test Battery was the selection instrument used by the Navy when the study was conducted.)

The current study has additional implications for developing a screening instrument for use when and if women no longer are required to have a high school degree for acceptance into the Navy. That is, the study suggests that the most valuable items for predicting attrition will probably come from the mental health, occupational need, enlistment motivation, and personal history areas (see Tables 2, 3, and 4).

REFERENCES

Campbell, D. P. *Handbook for the Strong vocational interest blank.* Stanford: Stanford University Press, 1971.

Craighill, M. D. Psychiatric aspects of women serving in the Army. *American Journal of Psychiatry*, 1947, *104*, 226-230.

Friedman, J. A modified screening questionnaire for service women. *U. S. Armed Forces Medical Journal*, 1956, *7*, 81-84.

Guilford, J. P., & Fruchter, B. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1978.

Hall, L. G. *Hall occupational orientation inventory*. Chicago: Follett Educational Corporation, 1971.

Horner, M. Fail: Bright women. *Psychology Today*, November 1969, 36-41.

Lipman-Blumen, J. How ideology shapes women's lives. *Scientific American*, January 1972, 34-42.

Lockman, R. F., & Gordon, P. L. *A revised SCREEN model for recruit selection and recruitment planning* (CRC 338). Arlington, VA, Center for Naval Analyses, 1977.

Plag, J. A., & Goffman, J. M. A formula for predicting effectiveness in the Navy from characteristics of high school students. *Psychology in the Schools*, Vol. III, No. 3, July 1966, 216-221.

Porter, L. W., & Steers, R. M. Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 1973, *80*, 151-176.

Sands, W. A. *Screening male applicants for Navy enlistment* (NPRDC TR 77-34). San Diego, CA: Navy Personnel Research and Development Center, 1977.

Schuckit, M. A., & Gunderson, E. K. E. *Psychiatric incidence rates of Navy women: Implications for an all-volunteer force* (Unpublished Report). San Diego, CA: Navy Medical Neuropsychiatric Research Unit, 1971.

REFERENCE NOTES

1. Statement of Vice Admiral James D. Watkins, Chief of Naval Personnel and Deputy Chief of Naval Operations for Manpower before the Subcommittee on Military Personnel of the House Armed Services Committee on Navy Proposal to Amend 10 USC 6015 (H. R. 6431) 21 March 1968.
2. Thomas, P. J. Impact of Pregnancy on the Navy. Briefing presented to the Chief of Naval Personnel, January 18, 1978.

Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Description of Leader Behavior

Major Jerome Adams Ph.D.

and

Jack M. Hicks Ph.D.

Prepared for presentation at the 1978 Military Testing Association Conference at the U.S. Coast Guard Institute, Oklahoma City, Oklahoma 30 October thru 2 November 1978

Running Head: Leader Behavior

* This paper represents the views of the authors and not the official position of the U.S. Military Academy, the U.S. Army, or any other governmental agency.

** The authors would like to thank Dr. Robert Priest and LTC Howard T. Prince II for their help in various phases of this research.

Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Description of Leader Behavior

ABSTRACT

In this paper the authors examine the relationship between male and female leaders description of their own behavior and the followers description of the leader's behavior in traditionally male-oriented leadership positions.

The data were collected as part of a larger research effort to assess how women are being assimilated into the Corps of Cadets at West Point, and how the women are being trained to become effective Army leaders.

During the summer of 1978, women cadets in the graduating class of 1980 were assigned for the first time into non-traditional platoon leader roles in predominantly male subordinate units. Both male and female platoon leaders were asked to describe their behavior using the Leadership Opinion Questionnaire (Fleishman, 1960). Two composite scores Consideration and Structure were the dimensions of leadership behavior. Subordinates in the platoons were asked to describe their leader's behavior on the same two dimensions, Consideration and Structure.

The results were interpreted in terms of three major issues: (1) the importance of sex roles as a leadership variable; (2) the leader perceptions of what performance behaviors are more important, Consideration versus Structure, and (3) the subordinates perceptions of what performance behaviors are important in a platoon leader's role.

INTRODUCTION

The concern about how well women can perform in non-traditional leadership roles has been a salient issue in the military particularly with the admission of women as cadets in the service academies. As military planners and researchers began to prepare programs for the development of women as future Army leaders, little empirical research was available in academic resources from which they could draw. Stogdill completed a comprehensive review of leadership research in 1974; however, sex roles and leadership were not systematically addressed. Terborg (1977) prepared a review of the literature on women in management roles. Some studies prior to 1975 suggest that there appears to be a bias in psychology for researchers to study males rather than females or both sexes (see Holmes and Jorgensen 1971; Dan and Beekman, 1972). Thus, military researchers and decision makers need to be cautioned about the generalizability of conclusions drawn from male-based research. Bender (1978) suggests that it remains unclear if social psychological literature on leadership is applicable for women as leaders.

This paper reports the results of a portion of a longitudinal research program to assess how women are being assimilated into the Corps of Cadets at West Point, and how effective the women are being trained to become effective Army officers.

RATIONALE OF THE STUDY

On October 7, 1975 President Ford signed into law Public Law 94-106, an amendment to which authorized women's admissions to the service academies, including West Point. As a result the academy developed operational plans for the admission of women as cadets.

Four phases of the program, later titled Project Athena, were planned:

- Preadmission phase to prepare cadets and the military community for the arrival of women (Vitters and Kinzer 1978).
- Integration phase which included careful documentation of how women were being integrated into the Corps of Cadets (see Vitters and Kinzer 1977, and Vitters, 1978).

- The Assimilation phase which studies how well women are being fully assimilated into the Corps of Cadets.
- The Graduate Assessment phase which will study how well women are performing their roles as officers.

The first two phases of Project Athena have been completed. The latter two are continuing to be designed and studied.

DESIGN

The design of this study involved five cadet companies where women were assigned into non-traditional roles as platoon leaders for the first time. The platoon leadership positions were for a four week interval after which a leadership change would occur. Women platoon leaders were assigned to both the first and second changeover detail.

At the end of the summer training, all platoon leaders were asked to describe their leadership behavior using Fleishman's Leadership Opinion Questionnaire. At a separate location, the subordinates were assembled to prepare peer ratings. During this time, the subordinates were also asked to describe the behavior of the platoon leaders of each detail using the same dimensions of Consideration and Structure.

Because there were only five women assigned in the non-traditional role as platoon leaders, a matched pair of five men from the same units on alternate details was used (see slide 1). Thus, the subordinates rated both the male and the female leader of the same platoon. The independent variables tested in the design were:

Cadet Companies*
Details within Companies (nested)
Platoon Leader Sex

The dependent measures used were:

Scores on the dimension of Consideration
(Welfare of subordinates)
Scores on the dimension of Structure
(Ability to get the task done)

* The company designations 1 thru 5 are arbitrary to protect the anonymity of the male and female leader participants.

FINDINGS:

In terms of differences between how male and female leaders describe their own behavior, there were no significant differences. That is, there was no significant difference between male and female platoon leaders in how they described themselves on the dimensions of consideration or structure. The authors conclude that the sample of only ten leaders was too small to note any sensitive differences between leaders on either of the criterion dimensions.

In the analyses where the subordinates described the leadership behavior of their leaders, statistically significant effects were noted. When the subordinates used Consideration as the dependent variable a leader sex main effect was noted (see slide 2). The slide shows that the platoon members perceived different behaviors on the part of male and female leaders with regard to the leader's concern for the welfare of the members.

However, because the significance tests do not provide any information about the pattern of effects, a multiple classification analysis was conducted to determine which sex provided more concern for subordinates (Consideration). The results of this analysis are presented in slide 3. The deviation from eta indicated in the LEADERSEX variables reveals that it is the females who are the leaders whom subordinates believe as having more concern for the welfare of the troops.

In the analyses where subordinates were asked to describe the leader behavior of their platoon leaders on Structure (Task Accomplishment) there were no main effects due to LEADERSEX. It is the authors' belief that the subordinates described their platoon leaders as equally capable of getting the task or mission accomplished. The multiple classification analysis revealed no significant difference between LEADERSEX for the Structure dimension (e.g., deviation eta for males -0.41 and 0.43 for females).

DISCUSSION:

The results reported in this study are part of a larger program which is trying to assess how well women are assimilating into the Corps of Cadets. Part of the assessment of full assimilation requires us to examine how well women are objectively performing in new, non-traditional roles as leaders and what the perceptions are about the women leaders' performance.

The data in this study indicates that the leaders themselves do not report any difference in how they see their platoon leader roles. This may well be an artifact in the methodology of too small a sample -- 10 leaders.

The more promising results indicate that subordinates do see male and female leader differences. Women are reported to be more sensitive to the welfare of subordinates Perhaps one may associate a priori the feminine communal values: sympathy, sensitivity and consideration as behaviors one may expect to typically find in women leaders (see Spence and Helmreich,1974). It is important to note that these behaviors are important for a leader -- especially one who will be expected to lead in an Army that requires the integrated services of both men and women.

Subordinates also reported no difference in leader behaviors between male and female platoon leaders in their activities to accomplish the mission (Structure). In this study, the authors are encouraged to find no statistically significant differences due to sex. Should men have higher subordinate scores on this dimension, one could possibly infer that there men were more inclined to get the job done than women.

The issues and concerns of how women are performing in new non-traditional roles will continue to be studied. Objective performance measures of how well women have performed in these roles is still being analyzed. Finally, comparisons of male superiors attitudes toward women in the Army and the superior's evaluations of men and women's performance is also being analyzed to see if any sex bias in evaluation of women leaders is unique to those male superiors with traditional beliefs.

Cadet Company	Leader Sex	Training Detail
1	Female	First Four Weeks Training
	Male	Second Four Weeks Training
2	Female	First Four Weeks Training
	Male	Second Four Weeks Training
3	Female	First Four Weeks Training
	Male	Second Four Weeks Training
4	Male	First Four Weeks Training
	Female	Second Four Weeks Training
5	Male	First Four Weeks Training
	Female	Second Four Weeks Training

* A sixth company was originally planned in the design however, the female who was designated to be the platoon leader voluntarily resigned and the orthogonal block of 3 women first detail 3 women second detail was lost.

SLIDE 1 Independent Variables: Company
Training Detail
Leader Sex

*HIERARCHIAL ANOVA: CRITERION (CONSIDERATION)

SOURCE	MEAN SQUARE	F	SIGNIFICANCE OF F
MAIN EFFECTS	170.30	2.46	.025
LEADERSEX	786.97	11.36	.001
DETAIL	8.43	0.12	.999
COMPANY	56.61	0.82	.999
2 WAY INTERACTIONS			
LEADERSEX COMPANY	151.96	2.19	0.088
EXPLAINED	164.19	2.37	0.014
RESIDUAL	69.27		

SLIDE 2 Leader Sex Main Effect for Subordinates description of leader behavior of Consideration

* Hierarchial approach (option 10) invokes the stepdown procedure. The sum of squares associated with the main effect for the first variable is not adjusted for any other variables. The sum of squares for the main effect for the second variable considered is adjusted only for the first variable, and so on (See Nie et.al., 1970)

MULTIPLE CLASSIFICATION ANALYSIS

VARIABLE & CATEGORY	UNADJUSTED DEV'N	ETA	ADJUSTED FOR INDEPTNDENT VARIABLES DEV'N	ETA
LEADERSEX				
1 MALE	-1.70		-1.71	
2 FEMALE	1.80		1.80	
		0.21		0.21
DETAIL		0.03		0.02
COMPANY		0.11		0.11

SLIDE 3 Multiple Classification Analyses

REFERENCES

Bender, L.S. Women As Leaders. Unpublished Dissertation: SUNY Buffalo, 1978.

Dan, A. and Beekman, S. Male Versus Female Representation in Psychological Research. American Psychologist. 1972, 27.

Fleishman, E. A. Manual for the Leadership Opinion Questionnaire. Chicago: Science Research Associates, 1960.

Holmes, D. S. and Jorgensen, B. W. Do Personality and Social Psychologists Study Men More than Women? Representative Research in Social Psychology. 1971, 2.

Nie, N. H. et. al. SPSS. 2nd ed., New York: McGraw Hill 1970.

Spence, J. T. and Helmreich,R. The Attitudes Toward Women Scale, Journal Supplement Abstract Service. 1972, 2.

Stogdill, R. M. Handbook of Leadership. New York: Free Press, 1974.

Ternborg,J. R. Women in Management: A Research View, Journal of Applied Psychology. 1977, 62.

Vitters, A. G. and Kinzer, N. S. Women at West Point: Change Within Tradition, Military Review. April, 1978.

Vitters, A. G. and Kinzer, N. S. Report of the Admission of Women to the US Military Academy (Project Athena I), USMA, New York, Sept., 1977.

Vitters, A. G. Report of the Admission of Women to the US Military Academy (Project Athena II). USMA New York, June 1978.

Female Utilization in Non-Traditional Areas

by

Joseph A. Bergmann
and
Raymond E. Christal

Air Force Human Resources Laboratory
Brooks AFB, Texas

INTRODUCTION

As current Air Force policy has opened many traditionally all-male specialties to women, it has become increasingly important to the Air Force to have detailed management information concerning how females are actually being utilized on the job. The Occupation and Manpower Research Division of the Air Force Human Resources Laboratory is devising methods of providing sufficient information to address present management questions regarding female utilization and identifying problems which, thus far, may not have received management's attention.

This report presents results of a probe study of female aircraft mechanics and outlines what we hope to accomplish in our follow-on in-depth analyses of the area. The probe study involved analysis of on-hand data collected during a routine occupational survey conducted by the Air Force Occupational Measurement Center (AFOMC).

METHOD

Survey Instrument

The job inventory used in the survey was very comprehensive, consisting of 977 task statements organized under 23 duties. It was administered between May and September 1976, and usable data were received from 5825 males and 206 females.

Analysis Sample

All of the women in our survey sample had been in the Air Force 43 or fewer months. This meant we could not simply compare male and female first-termers without further sample selection and matching. Furthermore, if we had restricted our samples to those who had been on board between 8 and 43 months, there would have been a 3-month difference in the average Total Active Federal Military Service (TAFMS) for males and females. In order to do direct comparative analyses, we matched the two samples on TAFMS case by case.

The final sample analyzed in our probe study is shown in Table 1. Our intent was to select males for a perfect month-by-month match with females on a five-to-one basis. We later had to discard one female for missing data, but this did not have a significant impact on the equivalence of the matched sample.

Data Quality Control Checks

In reviewing the survey returns, we noticed a number of individuals claiming to be females who had distinctively male names. We therefore matched the entire sample against the Air Force Uniform Airman Record file to eliminate possible errors in sex identification. We discovered that approximately one-half of one percent of the individuals surveyed made an error in identifying their own sex. This is not a very high error rate--only one in every two hundred subjects. However, in a sample of 6000 males and 200 females, this leads to an intolerable error in identification of the female subsample. In such an instance, 29 (or 12.7%) of the 229 identified as females would actually be males.

RESULTS

Job-Type Analysis

A very large number of job types and job-type clusters were identified using the Comprehensive Occupational Data Analysis Programs (CODAP) system. However, for simplicity, all of these could be classified as being either hard-core maintenance jobs or support jobs. Representative job types in these two categories are shown in Table 2. The title "Crew Chief" is somewhat of a misnomer. Crew chiefs are not supervisors; they are the flight line mechanics who perform primary aircraft maintenance tasks. Note that 59% of the males and 44% of the females in our sample were classified as crew chiefs. Some differences in job assignment as a function of sex is apparent from data in this table. Only 5.6% of the males were working in support jobs, while 26.2% of the females were working in jobs classified in this category.

Information in Table 3 suggests that during the first 43 months, there is a movement of individuals from maintenance to support jobs. However, this flow appears to be much larger for females than males. This implied difference is shown graphically in Figure 1.

It can be seen in Table 4 that women in maintenance jobs find their work at least as interesting as men and express at least equal intent to reenlist. They report their talents as being slightly less well utilized.

Data in Table 5 suggest that there are differences in the work performed by men and women working in support job types. Women spend more of their time performing tasks which can be classified as administrative or clerical in nature. However, there appears to be little difference in the nature of tasks performed by males and females working in maintenance jobs. Notice the small sex differences in time spent by male and female maintenance personnel on tasks classified as being "heavy" or "dirty." Also note that work performed by women in the support area is rated by supervisors as being more difficult than that performed by men either in the maintenance or support job areas.

Table 6 displays some of the differences in the duties performed by men and women in support jobs. Women spend more time maintaining records and forms. However, they also spend more time organizing, planning, directing, and implementing--which are duties normally performed by second-term personnel. The differences reported in this table are not highly stable because of the small numbers of cases involved.

The data in Table 7 reflect differences in the duties performed by men and women working in maintenance job types. This information should be fairly stable, since approximately 74% of the women and 94% of the men are working in these jobs. It appears that differences in the utilization of men and women in maintenance jobs are very small.

In one subanalysis, we identified 111 tasks which were performed by men in the sample, but not by women. However, you can see from Table 8 that not one of these tasks was being performed by as many as three percent of the men, and approximately one-half of them were being performed by less than one percent of the men. Inspection of these tasks led to the conclusion that women could and probably do perform them, but the small sample simply failed to pick up such cases.

Table 9 sums up one of the major findings in the probe study. The correlation of time spent on tasks by males with that of females is .97 in the maintenance job types. Even more striking is the correlation of .99 between the percent of males and females performing various tasks. It appears then that there is very little difference in the work performed by males and females in maintenance jobs. The relationship between the work performed by males and females in support jobs is considerably lower. As mentioned previously, this may have been due in part to unstable data as a function of sample size.

Analysis of Aptitude Distributions

We now turn to a second significant finding from the probe study. Table 10 reflects the aptitude requirement levels for entry into aircraft mechanic specialties for various time periods. Note that prior to June 1971 and since November 1975, a Mechanical Aptitude Index of 50 was required for entry. However, between these two periods, during which most of the individuals in the analysis sample came into the Air Force, applicants could qualify on either the Mechanical or Electronics Indexes at the 50th centile level.

Table 11 reflects gross differences in the Mechanical and Administrative Aptitude Indexes for the male and female members in the analysis sample. Notice that the mean mechanical AI for females was 39.2 which is considerably below the current 50th centile entry requirement level. Also, it was approximately one and three-quarters standard deviations below the mean score for males in our sample.

Presently there are approximately 2000 women working in the Aircraft Maintenance specialty. This makes possible one of the most definitive studies ever conducted on females working in a non-traditional area. The Personnel Research and Occupation and Manpower Research Divisions of AFHRL are presently completing details for a joint study of female aircraft mechanics.

The joint study will involve analysis of the complete input female sample to determine technical school success, attrition, retraining out of the specialty, and other factors relating to residualization of the group. For those still working in the specialty, we will evaluate their utilization patterns at the task level, survey their attitudes, evaluate their performance levels, and identify those who are moved from maintenance to support jobs to determine why. We will analyze task requirements for strength, stamina, and psychomotor skills and determine how such tasks are performed by members of both sexes. We hope to administer experimental tests of mechanical aptitude and validate them against current and future performance information. We will analyze promotion test scores and compare men and women supervisors. Finally, all cases will be followed throughout their careers to determine career patterns and attitude changes.

Our goal is to have sufficient information to address present and future management questions regarding female utilization in the Air Force.

The distributions of scores for the two subsamples in Table 12 are even more striking. Approximately 57% of the females in the probe study scored below the 50th centile and therefore could not presently qualify for entry into the Aircraft Maintenance career field. Yet, results from the probe study indicate that 75% of these women were working in mechanical job types and were performing essentially the same tasks as males. Furthermore, all of the women in the probe study had successfully completed technical training courses and had been working in the Aircraft Maintenance specialty for a number of months. However, this is a residualized group. We don't know how many male or female cohorts failed to graduate from school, and we don't know how many of them were retrained out of the specialty. Also, the present study is solely concerned with the time being spent on particular tasks by males and females. It does not address questions of the speed or quality of performance. All of these issues will be treated in the follow-on study. There is a growing body of evidence that mechanical aptitude tests which have historically been shown to be highly predictive of success for males may not be appropriate for females.

Table 13 lists some of the tests typically included in differential aptitude batteries such as the AQE and ASVAB. Automotive Information and Shop Information are primarily measures of mechanical experience. However, for the male population, it turns out that experience measures are good indicators of interest level and of ability to do well in subsequent mechanical training. Because of cultural differences, the same is not necessarily true for females. The Personnel Research Division of AFHRL is giving high priority to work on new mechanical aptitude measures which are appropriate for women.

CONCLUSIONS AND DISCUSSION

1. In traditionally all-male specialties recently opened to females, only a limited number of females can be expected. Caution should be exercised by those who may be extracting data on females from survey studies and accepting self-reported sex identification as being accurate.

2. Care should be taken to control for differences in lengths of service when comparing men and women in non-traditional specialties.

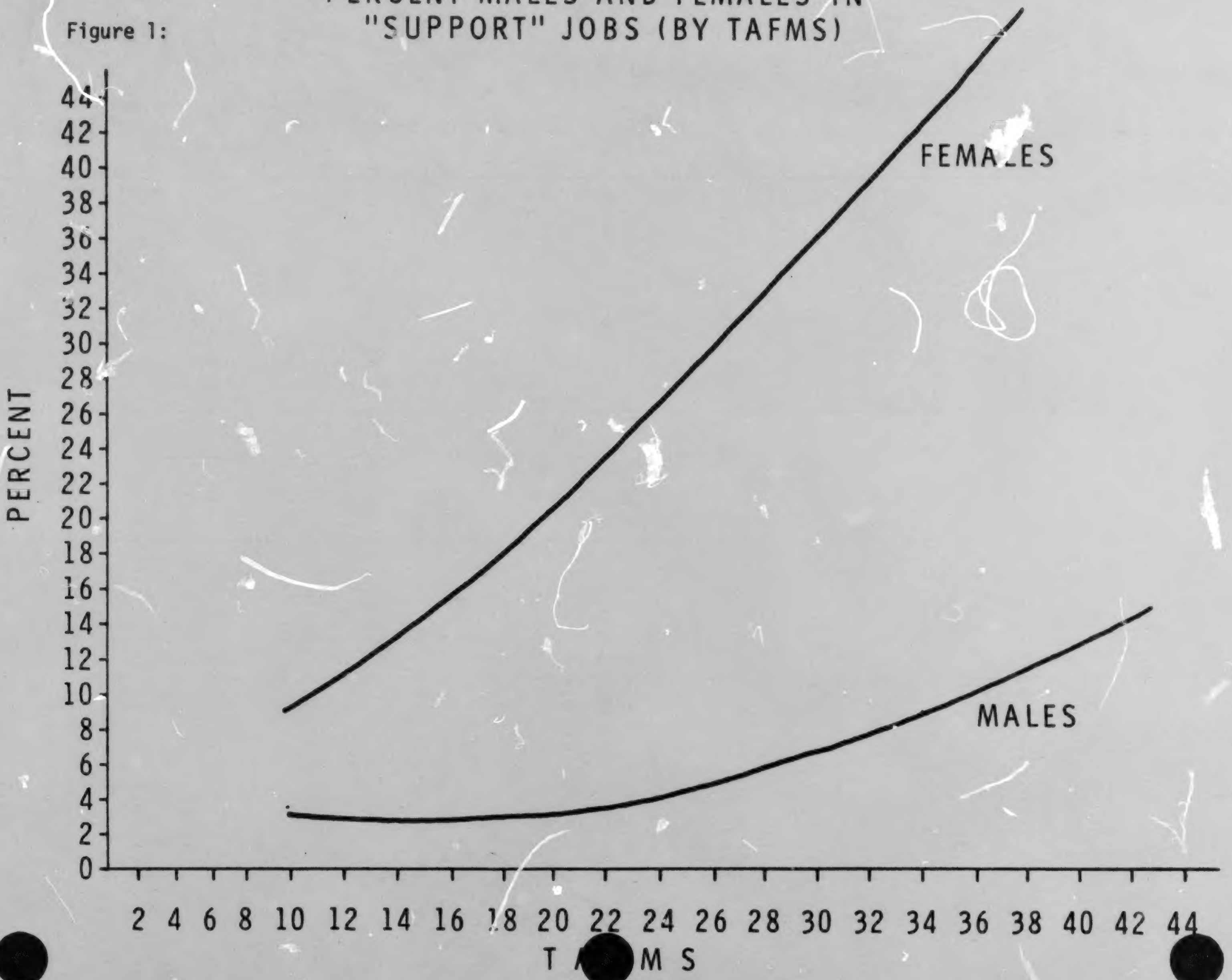
3. There is very little apparent difference in work performed by males and females in maintenance jobs within the 431X1 Aircraft Maintenance Specialty.

4. Females appear to migrate from maintenance to support jobs more rapidly and in higher proportion than males.

5. Mechanical aptitude tests highly predictive of success for males may not be appropriate for females in non-traditional specialties.

6. Additional research will investigate questions arising from the probe analyses.

Figure 1: PERCENT MALES AND FEMALES IN "SUPPORT" JOBS (BY TAFMS)



SELECTION OF SAMPLE FOR UTILIZATION OF WOMEN STUDY
AIRCRAFT MECHANICS 431X1 C, E, & F

8-43 MONTHS TAFMS SURVEY SAMPLE

		MALES	FEMALES
	N	1959	206
TAFMS	M	25.8	22.9
	S.D.	10.1	7.3

8-43 MONTHS ANALYSIS SAMPLE

		MALES	FEMALES
	N	1015	202
TAFMS	M	22.9	22.9
	S.D.	7.2	7.3

Table 2:

DISTRIBUTION OF MALES AND FEMALES IN SUPPORT AND MAINTENANCE JOBS

SUPPORT JOBS	MAINTENANCE JOBS
Tech Orders	Crew Chiefs*
Training	Inspection
Documentation	Special Maintenance
Scheduling	
Safety	
Job Control	
Deficiency Analysis	
Bench Stock	
CONTAINS:	CONTAINS:
5.6% of males 26.2% of females	94.4% of males 73.8% of females

$x^2 = 89.35$, df=1

$p < .001$

---

* 59% of all males and 44% of all females work in "Crew Chief" jobs

Table 3:

PERCENT MALES AND FEMALES IN "SUPPORT" JOB TYPES BY TAFMS

TAFMS	% OF MALES	% OF FEMALES
0-12	3.74	8.70
13-24	3.69	17.92
25-43	8.65	43.84
TOTAL	5.62	26.24

Table 4:

JOB ATTITUDES BY SEX AND JOB-TYPE CLASS 431X1 UTILIZATION OF WOMEN STUDY *

	SUPPORT		MAINTENANCE	
ATTITUDE **	MEAN MALE	MEAN FEMALE	MEAN MALE	MEAN FEMALE
Job Interest	4.53	4.36	4.64	4.72
Utilization of Talent	3.18	2.75	3.21	3.00
Utilization of Training	2.70	2.52	3.53	3.54
Reenlistment Intent	1.98	2.27	2.30	2.36

* S. D. About 1.3 - 1.6 for Attitude Variables; About 1.00 for Reenlistment Intent, Which is Defined As Follows:
1 = No; 2 = Uncertain, Probably No; 3 = Uncertain, Probably Yes; 4 = Yes

**No significant differences in t-tests between any male/female pairs.

Table 5:

PERCENT TIME SPENT ON VARIOUS CLASSES OF TASKS BY MEN AND WOMEN IN "SUPPORT" vs "MAINTENANCE" JOB TYPES

CLASS OF TASKS	SUPPORT		MAINTENANCE	
	MALE	FEMALE	MALE	FEMALE
Clerical	26.4	46.7	5.6	6.8
Heavy Maintenance	2.5	.3	8.6	8.3
Light Maintenance	5.6	1.2	14.6	13.9
"Dirty" Maintenance Tasks	.8	.6	8.6	8.3
Inspect, Check, Troubleshoot	8.6	3.2	37.2	37.2
Other (Support, Non-Clerical)	56.0	48.1	25.3	25.4
AVG # TASKS PERFORMED	28.1	18.8 **	157.9	141.1 **
AVG TASK DIFF. PER UNIT TIME	4.4	5.0 **	4.4	4.3

**p < .01 t-tests were not computed for class of task categories.

Table 6:

PERCENT TIME ON VARIOUS DUTIES FOR MALE AND FEMALE PERSONNEL WORKING IN SUPPORT JOB TYPES

DUTY	% TIME MALE	% TIME FEMALE	% TIME DIFFERENCE
F Performing Supply Functions	20.25	20.22	0.03
E Maintaining Forms and Records	17.34	28.15	-10.81
P Maintaining 780 Equipment	15.29	2.70	12.59
O Maintaining Non-Powered AGE Equipment	10.08	0.40	9.68
A Organizing and Planning	9.24	19.78	-10.54
G Performing General Aircraft Maintenance	6.70	3.30	3.40
B Directing and Implementing	6.42	12.65	- 6.26
D Training	4.94	4.60	0.34
C Inspecting and Evaluating	4.74	4.75	- 0.01
H Ground Handling of Aircraft	3.67	1.55	2.12
SUBTOTAL	98.67	98.10	

Table 7:

PERCENT TIME ON VARIOUS DUTIES FOR MALE AND FEMALE PERSONNEL WORKING IN MAINTENANCE JOB TYPES

DUTY	% TIME MALE	% TIME FEMALE	% TIME DIFFERENCE
G Performing General Aircraft Maintenance	22.61	22.53	0.08
H Performing Ground Handling of Aircraft	20.03	21.63	-1.60
I Maintaining Landing Gear Systems	11.24	9.88	1.36
M Maintaining Electrical Systems	6.32	7.27	-0.95
K Maintaining Flight Control Systems	5.19	4.56	0.63
Q Performing General Engine Maintenance	5.05	4.43	0.62
L Maintaining Pneudraulic Systems	5.00	5.37	-0.37
E Maintaining Forms and Records	4.79	5.55	-0.76
O Maintaining Non-Powered AGE Equipment	4.45	3.80	0.65
N Maintaining Fuel Systems	4.17	4.04	0.13
J Maintaining Utility Systems	3.23	3.14	0.09
F Performing Supply Functions	2.46	2.51	-0.05
C Inspecting and Evaluating	1.54	1.41	0.13
SUBTOTALS	96.08	96.12	

Table 8:

PERCENT OF MALES PERFORMING TASKS NOT BEING PERFORMED BY FEMALES

PERCENT OF MALES PERFORMING	NUMBER OF TASKS NOT PERFORMED BY FEMALES
3% or more	0
2.0 - 2.9	6
1.5 - 1.9	9
1.0 - 1.4	41
less than 1%	55
TOTAL	111 *

494

---

* of 977 tasks in the inventory

Table 9:

CORRELATIONS BETWEEN MALE AND FEMALE WORK IN 431X1 C, E, & F

VARIABLES	CORRELATION MALE VS FEMALE
% TIME SPENT ON TASKS IN MAINTENANCE JOB TYPES	.97 **
% PERFORMING TASKS IN MAINTENANCE JOB TYPES	.99 **
% TIME SPENT ON TASKS IN SUPPORT JOB TYPES	.69 **
% PERFORMING TASKS IN SUPPORT JOB TYPES	.58 **

**All correlations greater than 0 at .01 level.

Table 10:

APTITUDE REQUIREMENTS FOR ENTRY INTO 431X1

TIME PERIOD	REQUIREMENT LEVELS
Prior to June 1971	Mechanical 50
June 1971 - November 1975	Mechanical or Electronic 50
Since November 1975	Mechanical 50

Table 11:

APTITUDE INFORMATION FOR ANALYSIS SAMPLE
431X1 UTILIZATION OF WOMEN STUDY

APTITUDE COMPOSITES	MEAN MALE	MEAN FEMALE	t
Mechanical	68.9	39.2	21.55**
Electronic	64.5	60.7	3.90**
General	60.7	69.9	7.46**
Administrative	47.6	63.4	10.79**

$^{**}p < .01$

Table 12:

DISTRIBUTION OF APTITUDE SCORES FOR ANALYSIS SAMPLE BY SEX

	MECHANICAL AI		ELECTRONIC AI	
SCORE	MALE	FEMALE	MALE	FEMALE
0	0	0	0	0
5	0	4	1	0
10	0	9	2	1
15	2	16	1	0
20	1	7	4	0
25	7	33	6	0
30	7	14	7	0
35	9	12	11	3
40	11	12	57	3
45	10	5	42	1
50	96	23	82	36
55	138	27	115	41
60	129	22	131	42
65	87	7	114	26
70	76	2	105	14
75	76	3	83	14
80	65	1	74	6
85	85	0	67	3
90	114	0	47	5
95	69	0	33	2
MEAN	68.93	39.16	64.45	60.66
SD	16.29	17.96	15.86	11.66

Table 13:

MECHANICAL SUBTESTS IN AQE/ASVAB

Mechanical Principles

Automotive Information

Shop Information

SECTION 4

GENERAL

501

Strain by prolonged duty hours and problems as to mobility of soldiers - as seen by the Federal Armed Forces Association.

By Colonel H.E. Seuberlich.

I. Introduction.

The German Federal Armed Forces Association, founded in 1956 by 55 soldiers, today, as a top organisation with more than 230,000 members of all status groups and ranks, represents the professional and social interests of servicemen. Its highest authority is the General Meeting. From it the Federal Board receives its commission. The 10th General Meeting passed the programme for the next four years with 300 resolutions.

The individual resolutions reflect the manifold problems of the servicemen of the Federal Armed Forces in the late 70s. With their Comprehensive fundemental programme, they will be taking effect far into the next decade. Thus they are now already forming the image of the serviceman of the 80s.

Allow me to cite two crucial resolutions. One demands a stock-taking of the "personal and social situation in the Federal Armed Forces." Defence Minister Apel has already introduced this in the meantime. The second resolution which I want to mention outlines a general defence concept for the Federal Republic of Germany, it includes, inter alia, the demand for compulsary service in which women should also be included.
In this consideration the goals are:

- the improvement of conscriptional equality
- securing the necessary personel in the second half of the 80s if the rise in the number of those liable to military service begins to sink as a result of the structure of age groups in the Federal Republic of Germany.
- increased motivation for the military service.

These are indications of the future. Both should contribute to the abolition of the disturbance of the equilibrium which has arisen in recent years, and which have formed a focal point for measures and budgetary means with the change in armament and modernisation of the German Federal Armed Forces with arms systems of the future. The Federal Armed Forces Association emphasizes that despite technical perfection, the human factor should not be rejected, because even the technically most advanced arms system in effect is only as effective as the serviceman who operates it, and as his sincere willingness to do his best. This willingness will be encouraged or impaired by:

- his social situation
- the strain imposed by his daily service
- his job satisfaction
- many influences from the outside world, especially from his own family.

His committment in an emergency will be largely conditioned by the knowledge of how his family and the civil population will be protected from attack.

II. Job evaluation.

1. The factual situation of job evaluation in the public service of the Federal Republic of Germany.

Since 1975 the basic principle for the function orientated pay has been contained in the Federal Law of Payment. According to this the level of payment of civil servants, judges, and soldiers should be determined according to the importance of their fufilled functions. Therefore their functions should be properly assessed according to their requirements and the appropriate ranks should be assigned. The aim in this respect is just payment in the civil service. That requires a standardized scale of assessment for all departments of the public service.

The Federal Ministry of the Interior is responsible for this. It is working on the development of a relevant REFA system. A first file of characteristics, for the civil service only, has already been tested in 281 jobs. A second edition is being tested now until next year on a larger scale.

2. Conception and cooperation of the German Federal Armed Forces Association.

The hitherto existing files of characteristics do not yet record the service characteristics of servicemen with their manifold special demands and stresses, or do so insufficiently. Therefore it is not clear either how they should be evaluated. At its 10th General Meeting the German Federal Armed Forces Association dealt with these problems in detail and passed a draft conception. From this a few basic principles :

- The special requirements and stresses on servicemen as opposed to other sections of the public service, are to be evaluated according to standardized scales.
- the claim on servicemen's time, the frequent separation from their families and the frequent change of duty must also be considered.

- It must be possible to consider fairly any probationary periods or experience necessary

- A comprehensive analysis of the requirements of all posts in the Armed Forces will be necessary for this evaluation.
- This requires a job specification which makes the determination of a verifiable job evaluation possible. This will be served only by a file of characteristics with the typical characteristics of the activities in the armed services for a standardized listing. This file is to be coordinated with the files of the other sections of the public service.

The army chairman, Colonel Seuberlich, has been commissioned with the representation of these principles within the scope of the present activities of the Ministry of the Interior.

3. The activities of the Federal Ministry of Defence.

The comprehensive analyses of the requirements for all posts in the armed forces commissioned by the Federal Armed Forces Association serve the projects "function analyses of the personel structure" undertaken by the Federal Armed Forces Association already known to you. They are based on the suggestions made by the Commission on Personel Structure which were also lectured on several times before the MTA. The Federal Ministry of the Interior is aiming for an agreement on the critical points with the Federal Ministry of Defence by January 1979, to the effect that also the servicemen in the next trial period can be included.

III. Strain by prolonged duty hours.

1. The evolution of working time regulations.

When the Federal Armed Forces were established in 1956 the official working hours were 48 hours per week. Within 15 years this was reduced to a 42-hour week. In 1974 the 40 hour week was even introduced into the public service. The Chancellor of the Federal Republic recognized this reduction with a 5% pay rise .Civil servants' overtime over and above this which can be measured and cannot be compensated for by free time, is reimbursed.

The reimbursement,which has been raised several times in recent years, at the moment amounts to between 9.50 DM and 18.50 DM per hour for civil servants of the various pay categories.

2. The situation of the serviceman.

For servicemen, however, neither the regulation of working hours nor the payment of overtime has up till now been planned. As they receive no other compensation, their social equilibrium is considerably disturbed as far as wages are concerned, as : according to statistics submitted by the German Federal Armed Forces Association to the Lower House of the Federal Government in 1978 soldiers work

- 60% regularly up to 50 hours a week
- 12% between 51 and 60 hours a week
- 5% over 60 hours a week

Manoeuvres and military excercises of on average 40 days per year are not included in these figures, although that would in most cases bring the hours up to about 80 per week.

A general regulation of working hours for servicemen cannot be coordinated with the necessary readiness for action of the armed forces with the present number of personel. Nevertheless this special stress must be entered in the file of characteristics for servicemen, in order to unequivocally record the disturbed social equilibrium , and to further the search for a possible solution. That such a solution exists is shown by the example of policemen , who are comparable to servicemen, whether it be the Federal Border Police or the police forces of the states of the Federal Republic of Germany, with a 40 hour week and overtime pay.

III. The ascertainment of normal duty hours for different groups.

Here it is neither a matter of the introduction of a 40 hour week for servicemen, nor of the creation of the basic requirements necessary so that servicemen could receive overtime pay. Both would be a gross misunderstanding. It would lead to a bureaucratized army and to the time clock serviceman, which the Federal Armed Forces Association decidedly rejects.

The Association acts rather from the assumption that in the modern armed forces the demands of time on the serviceman arising as a result of military excercises, manoeuvres and duties of various kinds vary greatly in the course of a year

They are, moreover, dependant on the different demands and situations in the individual branches of the armed forces, distributed among the SOLL personel. It is therefore the opinion of the Federal Armed Forces Association that these connections should be investigated in detail, and that subsequently the relevant conclusions should be drawn. These could be concerned with organisation, finances and personel. They should, though, be concerned with the "normal working hours " of whole units, and not those of the individual serviceman. Thus 6 to 10 large groups will be formed of servicemen exposed to similar or comparable stress, and which would be under consideration to find feasable and socially balanced justifiable solutions.

IV, Mobility

Mobility is one of the characteristics of service in the armed forces. For the serviceman it therefore entails the obligation to allow himself to be transferred at any time, to take part in training courses or to take on new duties. This characteristic of service in the armed forces , often linked with a change of base, affects about a quarter to a third of all professional servicemen and "Zeitsoldaten" (volunteers who sign up for a certain number of years) every year. Unmarried people take this more or less in their stride, but married servicemen are as a rule confronted with many problems. The Federal Armed Forces Association, inter alia, has investigatied them with their wives in two symposia, and the results can thus be summerized :

- children are the worst affected; they are uprooted from familiar surroundings, have to change schools, lose friends and other social ties. The slogan "if the father is transfered the child has to repeat a year at school" characterizes however only one aspect of the problem, when the child is not able to continue his education at a new school without any breaks or inconsistencies, as a result of the confusion created by the particularistic educational policy of Federal Germany.

Inner conflicts, possibly involving psychological damage,- whether it is the children or the wives who discover only after a matter of years that they are unable to cope with a constant change of address.

- For children the frequent change of address entails a lasting negative influence on their future careers; not counting the inferior educational opportunities for school-leavers in economically weak areas in which military bases are often situated.

- the wives are also forced to make sacrifices which are not asked of other women of our society. That opportunity of self-realization through a career is denied them which our women strive towards with increasing zeal as a result of their new self-confidence.

- Both the woman's share in the family's earning power and her own occupational and social security are reduced to a minimum.

- Financial sacrifices,- not counting those incurred by moving house,- are often caused by high rents, which even pay rises through promotion do not cover: and a transfer is not always coupled with promotion anyway.

Nevertheless, frequent transfers do not affect all servicemen equally. They vary according to the different ranks and duties.

V. Possibilities and limits on the conclusions for the consequences of strain by prolonged duty hours and mobility.

Both strain by prolonged duty hours and mobility contain characteristics of service in the armed forces. Both have one thing in common - they are not directly connected with a definite post and its demands. Therefore both exceed the systematics of an analysis of the requirements concerned with a post in the services and its proper assessment.

The situation is further complicated by the fact that both components have a certain, at times close correlation to one anther. For every new duty demands a period of vocational adjustment which usually also entails additional working hours. That means that the frequency of the change of post increases the strain by prolonged duty hours.

While "normal working hours" can be established according to organisational fields, strain as a result of mobility has to be established according to service ranks and duty or training groups.

General possibilities for the alleviation of the negative consequences of mobility could be found in :

- more generous compensation for expenses entailed by moving house.

-

507

- measures for the standardization of rents in the different bases.

- Assistance in the integration of families in new bases in all the different areas of life.

VI. Resumé and outlook

The phenomena of strain by prolonged duty hours and mobility are becoming noticably a problem for the armed forces, because they are no longer accepted by servicemen and their families as inevitable, but are compared with the working conditions of tohers.

Organisors, personel planners, those who draw up plans for trainig schemes and work timetables are therefore cooperating more and more closely with one another. With regard to general development in the working society, it is impossible in the long term to try to explain away social inequalities simply as the characteristics of work in the armed forces, without seriously endangering the necessary motivation.

In a file of characteristics for servicemen, therefore, strain by prolonged duty hours, and mobility must also be considered as factors of work science.

In its role as social early warning system the German Federal Armed Forces Association has for years been drawing attention to the complexity of these problems and in the future will also point out concrete possibilities for a solution.

Computer Assisted Reference Locator (CARL) System: An Overview[1]

by

William A. Sands

Acquisition and Initial Service Program
Navy Personnel Research and Development Center
San Diego, California 92152

The 20th Annual Military Testing Association Conference

Oklahoma City, Oklahoma

30 October - 3 November 1978

---

[1] The opinions or assertions contained herein are those of the writer and are not to be construed as official or reflecting the views of the Navy Department.

INTRODUCTION

Background

The problem which originally prompted the present author's interest in the field of information retrieval is humorously related by Redican (1973) in an article entitled "Reprints: File Before They Defile You," published in the American Psychologist. He states that:

> Several centuries ago, a philosopher, whose name now escapes me, is said to have come to an untimely end when his overloaded bookcase toppled over and buried him. Although the condition of most psychologist's bookshelves is undoubtedly not quite so lethal, many are burdened by sizable piles of reprints, manuscripts, journals, and books that await filing (July 1973, p. 625).

The creation, utilization, and maintenance of a reference retrieval system was identified as a professional problem for the field of psychology twenty-five years ago by Daniel and Louttit (1953). The magnitude of the problem can be appreciated by the recognition that the growth rate of scientific publications is exponential, with the number of references doubling every 13-15 years (Price, 1961, 1963). This rate of information production can be overwhelming to the individual researcher who attempts to keep up with the literature in a particular field. The initial attempt at imposing order on a growing personal reference library often involves filing documents alphabetically by the last name of the senior author. Inevitably there arises a situation where the researcher knows that the library contains a reference on a particular topic but cannot recall the author's name. Location of the reference requires a sequential search of the alphabetical file of source documents. With a small reference library, this sequential search process is inconvenient. As time passes and the size of the reference library grows, the sequential search strategy becomes increasingly burdensome. Therefore, some alternative information retrieval method is sought.

Information Retrieval Systems

Lancaster (1968), adopting a broad perspective, maintains that information retrieval encompasses all the activities from the initial acquisition and indexing of source documents to the search, retrieval, and delivery of the results of a query to the user. He further points out that an information retrieval system does not change the knowledge of the user on any subject. Rather, the system simply informs the user of the presence or absence of source documents on a particular topic and the location of all pertinent documents.

A wide variety of information retrieval systems is available (Bourne, 1963; Lancaster, 1968) including a manual system using ordinary index cards, an edge-punch card system employing thin rods to sort and retrieve information, and various computer-based systems. One manual

510

approach which has considerable merit is the "accession number coordinated system" described by de Alarcón (1969), and, specifically, a variation known as "Uniterm" introduced by Taube (1953) and discussed in the American Psychologist by Broadhurst (1962). The coordinate indexing method advocated by Broadhurst is described as follows:

> The procedure involves setting up a separate card file to index the collection of references, these being filed irrespective of author or content but merely according to a serial number assigned to each reference as it is added--technically the accession number. The classification of the reference is then done by selecting certain key words and underlining them ('tracing'). This is most conveniently done as the reference is read, and can be done on the card or the associated reprint if you have one, or in your copy of the journal, so long as the serial number given to either of the latter is the same as that shown on the reference card. These key words (or 'Uniterms') which have been created by the tracing procedure can be as many or as few as you like, depending on your interests and the relevance of the material to them. If an important classification does not occur in the text or reference, it can be added. The serial number of the reference is then transferred ('posted') to cards which merely bear the appropriate Uniterms as headings. Proceeding in this way generates a personal psychological vocabulary of Uniterms represented by cards each having the serial numbers of references on them. Retrieval of information then becomes simple. Consideration of any one Uniterm card will give the serial numbers of all the references which deal with the subject in question. For two subjects, take the two Uniterm cards, and compare them for coincidences of number. Such coincidences of number indicate references which deal with both subjects. Making a series of such cross matches of Uniterm cards will yield information about the collection of references in as broad or as fine a detail as is required. (1962, p. 137).

Thus the term coordinate indexing is quite descriptive. Each Uniterm can be considered as one of a set of classification coordinates. For any pair of Uniterms, the accession number(s) at which the pair intersects represents a reference(s) that has been classified as belonging under both Uniterms.

THE CARL SYSTEM

Design Objectives

The Computer Assisted Reference Locator (CARL) System is a computer-based information retrieval system which generally follows the coordinate indexing approach described above. Some of the objectives considered in designing the system were: (1) simplicity of reference query and

511

retrieval; (2) ease of system maintenance (additions, deletions, corrections, etc.); and (3) adaptability for alternative computer systems.

The first objective, query-retrieval simplicity, is primary. If a system is too complicated, troublesome or time-consuming, a researcher probably will avoid using it. From a practical standpoint, any information retrieval system which is not used is worthless, regardless of the creativity displayed in the system design or programming.

The second objective, maintenance simplicity, is also important. If the system is so complex that changes to the existing data files or the addition of new references to the system is a monumental task, the system probably would be expensive in terms of time, costs, or both. Obviously, this expense could constrain the utility of a reference retrieval system.

The third objective, adaptability, is a desirable characteristic as it will facilitate the adoption and use of the CARL System by researchers having access to a wide variety of computer systems with different operating systems, internal and external storage conventions, and input/output characteristics. The system design and the component source programs of the CARL System have been developed with this third objective in mind. The original version of the CARL System (described in this paper) is a sequential access system, as distinguished from a direct access system. Therefore, the data files which will be described below could be stored either on magnetic tape or on disk. Most computer systems include one or both of these storage media. The source programs which incorporate the processing logic of the CARL System are written in ASCII FORTRAN.[2] The use of FORTRAN should insure a wide degree of compatibility with different computers, as most systems support FORTRAN.[3]

Input Information

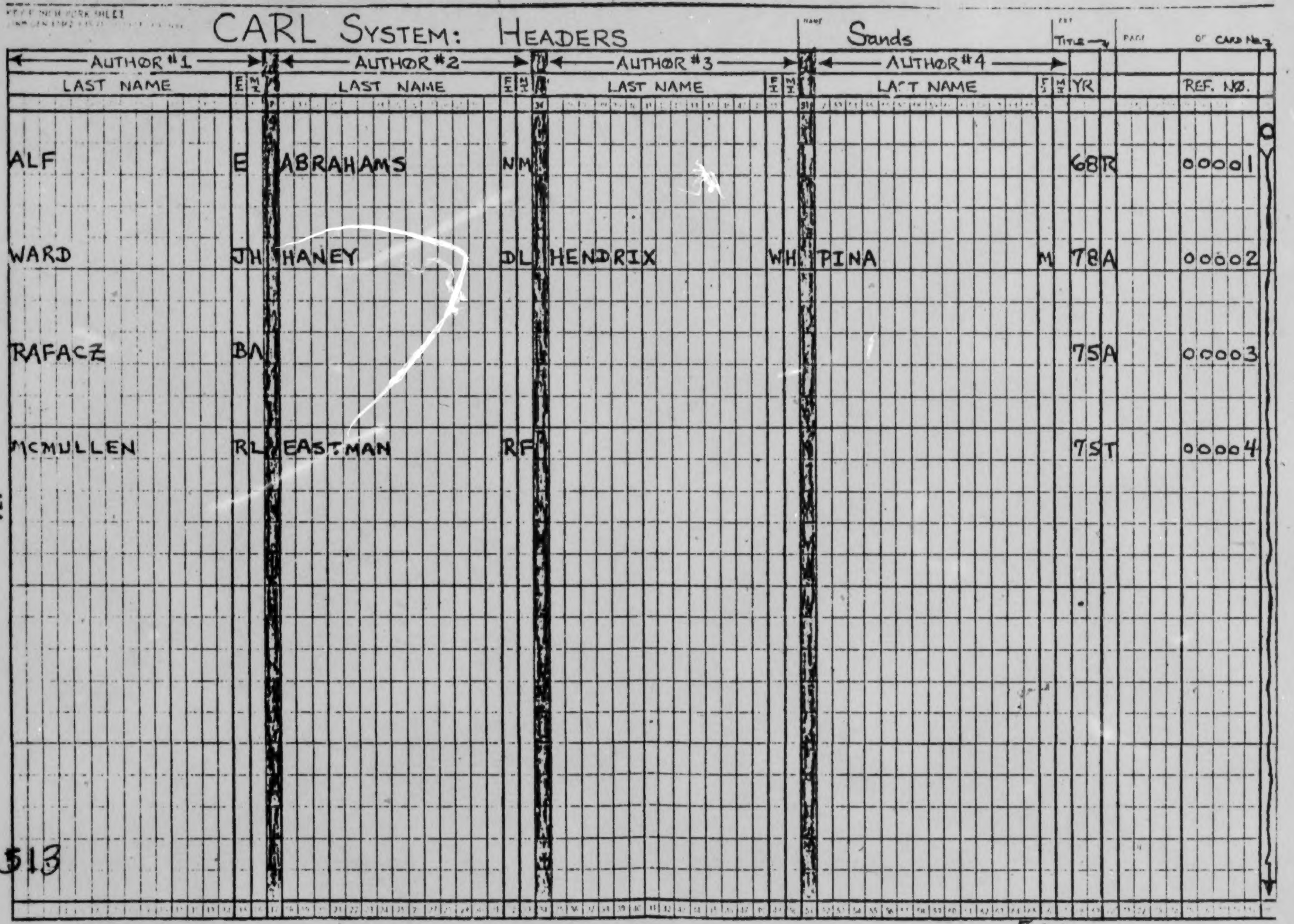
The information for each new reference is encoded for keypunching[4] on four different forms. The first form is for headers, as shown in Figure 1. The header card allows space for up to four authors (last name and first and middle initials), the year of publication, and the first letter of the reference title. In addition, there is space for a five-digit reference number and a one-digit card number. Zero is the card number for the header card. Each reference has only one header card, regardless of the actual number of authors. The header card is used for sorting operations in the CARL System.

---

[2] A version of the FORTRAN language which handles the full American Standard Code for Information Interchange character set.

[3] The computer system used in the development of the CARL System is a UNIVAC 1110, a general purpose, high performance, multiprocessor system employing the EXEC 8 executive system. This computer system is located at the Naval Ocean Systems Center in San Diego, California.

[4] Actually, there are no system constraints which require card input. The information could be entered via a computer terminal.



CARL SYSTEM: HEADERS

NAME Sands

TITLE →

PAGE OF CARD No.

AUTHOR #1 LAST NAME	F I	M I	AUTHOR #2 LAST NAME	F I	M I	AUTHOR #3 LAST NAME	F I	M I	AUTHOR #4 LAST NAME	F I	M I	YR		REF. NO.
ALF	E		ABRAHAMS	N	M							68	R	00001
WARD	J	H	HANEY	D	L	HENDRIX	W	H	PINA	M		78	A	00002
RAFACZ	B	N										75	A	00003
MCMULLEN	R	L	EASTMAN	R	F							75	T	00004

Figure 1. Header [illegible] Keypunching Form.

The second form is used for the text of the reference citation. There are no system constraints on the arrangement of information within this text card (excluding the reference number and card number). Figure 2 illustrates the conventions adopted by the present author. The first line of text begins in column one with the senior author's last name, first and middle initials. This is followed by the article title, the journal name, year, volume number, and page numbers. Note that the additional text cards differ from the first text card in that information begins in column four. This is done to improve readability of lists of references. The second example in this figure illustrates the encoding of a technical report published by one of the military personnel research laboratories. The authors and title are formatted as above, followed by the report type and number assigned. Next, the location and the name of the performing organization are specified, followed by the publication date. The text cards for a single reference are assigned card numbers from one to six, as needed.

The third type of form is for author information. As shown in Figure 3, the format allows for four authors per card. If there are more than four authors, a second author card is used. All author cards are assigned the number seven as the card number.

The last form used to encode input data is for keywords. As shown in Figure 4, up to four keywords are allowed per card. The number of keyword cards for a single reference is unlimited, but each keyword card is assigned eight as a card number. The only system constraint on keywords is a length of eighteen characters. The present author has developed a controlled vocabulary for use in assigning keywords to references. The choice of keywords for a reference is the most crucial aspect of encoding all input data. If many keywords having only a remote relationship to the reference are assigned, the reference will be retrieved frequently when it is not useful. On the other hand, if a reference is not assigned critical keywords, it will be missed in a reference search even though the material is pertinent to the user's needs. The best balance between these two considerations will depend upon the typical user of the system. The present author leans toward overinclusion (i.e., too many keywords), to insure that all pertinent references are identified in a search.

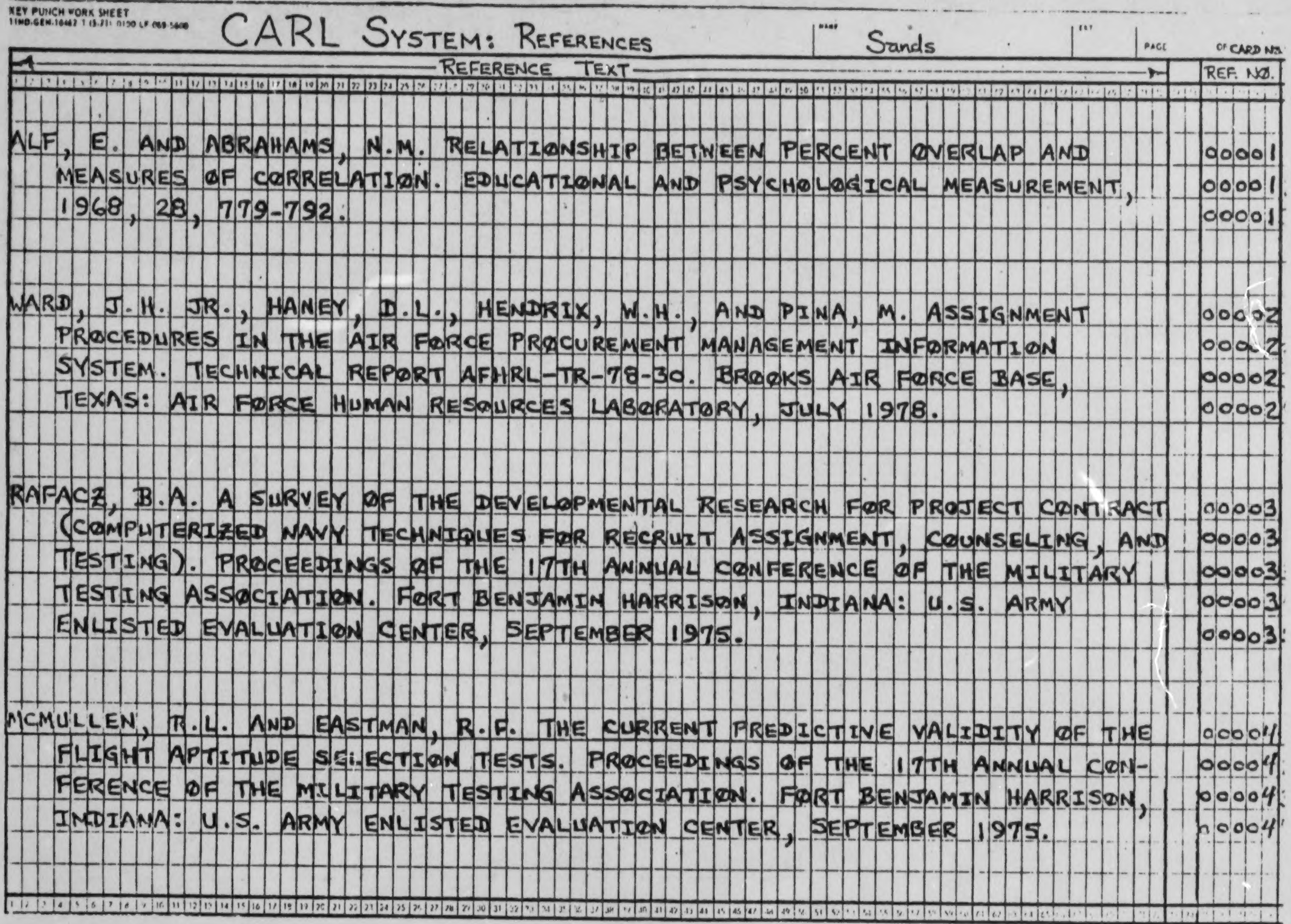
New References

The addition of new references to an existing library involves twelve steps:

1. Alphabetize a new set of source documents by author(s) and remove any duplicates.

2. Check new source documents against existing library to identify apparent duplicates.

3. Determine unique/duplicate status of potential duplicates and eliminate identified duplicates.



KEY PUNCH WORK SHEET
11ND-GEN-10462 1 (5-71) 0100 LF 059 5600

CARL SYSTEM: REFERENCES

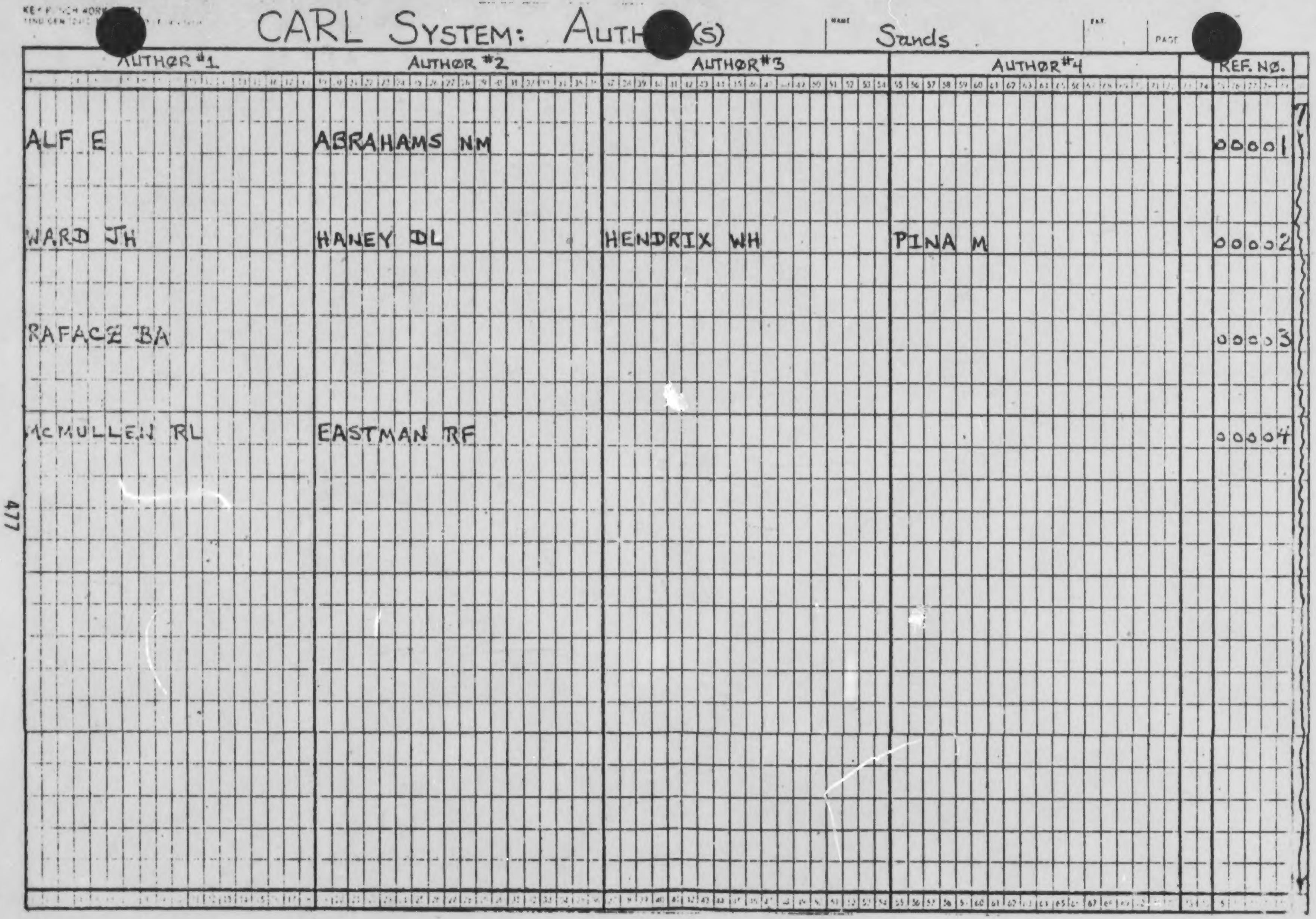
NAME Sands

PAGE

OF CARD NO.

REFERENCE TEXT	REF. NO.
ALF, E. AND ABRAHAMS, N.M. RELATIONSHIP BETWEEN PERCENT OVERLAP AND	00001
MEASURES OF CORRELATION. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT,	00001
1968, 28, 779-792.	00001
WARD, J.H. JR., HANEY, D.L., HENDRIX, W.H., AND PINA, M. ASSIGNMENT	00002
PROCEDURES IN THE AIR FORCE PROCUREMENT MANAGEMENT INFORMATION	00002
SYSTEM. TECHNICAL REPORT AFHRL-TR-78-30. BROOKS AIR FORCE BASE,	00002
TEXAS: AIR FORCE HUMAN RESOURCES LABORATORY, JULY 1978.	00002
RAFACZ, B.A. A SURVEY OF THE DEVELOPMENTAL RESEARCH FOR PROJECT CONTRACT	00003
(COMPUTERIZED NAVY TECHNIQUES FOR RECRUIT ASSIGNMENT, COUNSELING, AND	00003
TESTING). PROCEEDINGS OF THE 17TH ANNUAL CONFERENCE OF THE MILITARY	00003
TESTING ASSOCIATION. FORT BENJAMIN HARRISON, INDIANA: U.S. ARMY	00003
ENLISTED EVALUATION CENTER, SEPTEMBER 1975.	00003
MCMULLEN, R.L. AND EASTMAN, R.F. THE CURRENT PREDICTIVE VALIDITY OF THE	00004
FLIGHT APTITUDE SELECTION TESTS. PROCEEDINGS OF THE 17TH ANNUAL CON-	00004
FERENCE OF THE MILITARY TESTING ASSOCIATION. FORT BENJAMIN HARRISON,	00004
INDIANA: U.S. ARMY ENLISTED EVALUATION CENTER, SEPTEMBER 1975.	00004

Figure 2. Text C Keypunching Form.



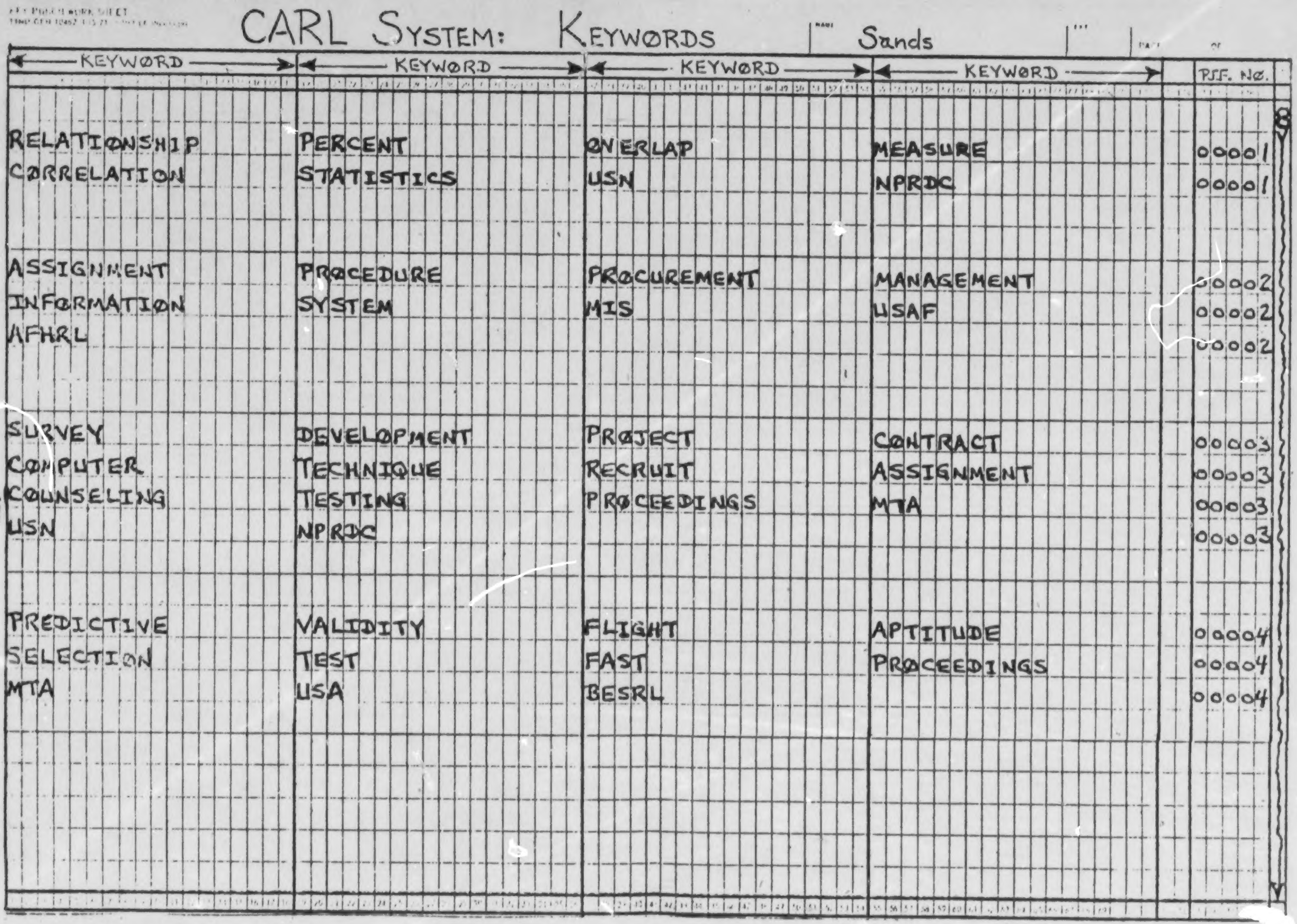
CARL SYSTEM: AUTH (S)

NAME Sands

PAGE

AUTHOR #1	AUTHOR #2	AUTHOR #3	AUTHOR #4	REF. NO.
ALF E	ABRAHAMS NM			00001
WARD JH	HANEY DL	HENDRIX WH	PINA M	00002
RAFACE BA				00003
MCMULLEN RL	EASTMAN RF			00004

Figure 3. Author Card Keypunching Form.



KEY PUNCH WORK SHEET

CARL SYSTEM: KEYWORDS

NAME Sands

PAGE OF

KEYWORD	KEYWORD	KEYWORD	KEYWORD	REF. NO.
RELATIONSHIP	PERCENT	OVERLAP	MEASURE	00001
CORRELATION	STATISTICS	USN	NPRDC	00001
ASSIGNMENT	PROCEDURE	PROCUREMENT	MANAGEMENT	00002
INFORMATION	SYSTEM	MIS	USAF	00002
AFHRL				00002
SURVEY	DEVELOPMENT	PROJECT	CONTRACT	00003
COMPUTER	TECHNIQUE	RECRUIT	ASSIGNMENT	00003
COUNSELING	TESTING	PROCEEDINGS	MTA	00003
USN	NPRDC			00003
PREDICTIVE	VALIDITY	FLIGHT	APTITUDE	00004
SELECTION	TEST	FAST	PROCEEDINGS	00004
MTA	USA	BESRL		00004

478

Figure 4. Keyword [illegible] Keypunching Form.

4. Code information on new documents onto keypunch forms:
    - A. Header form for Card #0
    - B. Text form for Cards #1-6
    - C. Author(s) form for Card #7
    - D. Keyword(s) form for Card #8

5. Keypunch new information (Cards #0-8) for each new source document.

6. Verify keypunching of new information.

7. Sort new card deck by document number and card number.

8. Run edit program on sorted deck.

9. Correct any problems identified by edit program.

10. Proofread edit program output:
    - A. Header
    - B. Text
    - C. Author(s)
    - D. Keyword(s)

11. Correct any problems identified by proofreading.

12. Input corrected card deck and update system.

Data Files

The CARL System data files may be divided into those primary files accessed by the system during normal processing and those backup files kept as insurance against disaster.

There are four primary data files in the CARL System.[5]

1. RDXXXX.--The raw data file which contains all the raw data from the header, text, author, and keyword cards for each reference (i.e., cards zero through eight), arranged sequentially by reference number and card number.

2. KAXXXX.--The keyword-author file which contains all of the keywords and authors with the associated reference numbers, arranged alphabetically by keyword/author.

---

[5] The XXXX portion of each data file name is a number indicating the highest reference number currently incorporated into the system. For example, RD0400. would indicate a data file based upon references 1-400, inclusive.

3. KDXXXX.--The keyword dictionary file containing all keywords employed in the existing library, arranged in alphabetical sequence.

4. ADXXXX.--The author dictionary file containing all the authors employed in the existing library, arranged in alphabetical sequence.

All four primary data files are backed up on magnetic tape. Specifically, the raw data file (RDXXXX.) and the keyword-author file (KAXXXX.) are kept on one reel and the two dictionary files (KDXXXX. and ADXXXX.) are kept on another reel. Finally, the original punched cards are saved so that the entire CARL System could be rebuilt if all the disk files and magnetic tapes were destroyed.

Computer Programs

The computer programs in the CARL System can be divided into two categories: (1) input preparation programs, and (2) system programs. As the name implies, the purpose of the input preparation programs is to clean up the input data prior to incorporating it into the CARL System. The EDIT program reads the punched card deck (or a disk file) containing all the raw data and prints it out in a format which facilitates visual editing. In addition, the program checks for the following problems: (1) cards missing or not in sequence within a reference, (2) reference numbers not in sequence, (3) missing reference numbers, (4) duplicate reference numbers, and (5) illegal reference numbers. Appropriate diagnostic messages are printed as error conditions are encountered. Finally, after processing all the input data, the number of errors flagged and the number of missing references are reported.

The duplicate identification program (DUPL) reads the raw data file (RDXXXX.) and strips off all zero cards. These header cards are then sorted on four fields in the following order of significance: (1) author, (2) publication date, (3) first letter of title, and (4) reference number. A sorted listing of the header cards is produced. Optional outputs include: (1) identification and listing of all potential duplicate references (i.e., references with identical header cards) and (2) a comparison of potential duplicate references with a stored list of apparent duplicates which have been previously identified as unique references and a listing of the remaining potential duplicates.

The recommended keyword dictionary program (RKWD) reads a punched card deck of keywords, arranges them in alphabetical order, and prints out a dictionary of recommended keywords. This dictionary is designed to provide initial guidance for indexers as a system is getting started and includes only a few references. Later, after a substantial number of references has been indexed and incorporated into the system, a dictionary of actual keywords will be used.

There are three systems programs: (1) a system creation program (BUILD), (2) a query-retrieval program (QUERY), and (3) a system maintenance program (CHANGE). The BUILD program is used only once (assuming the data files in the CARL System are not destroyed). As shown in Figure 5, the BUILD program reads the raw data either from the original punched card deck or a disk file containing the same information. The data are edited and if an abortive error condition is encountered a message is printed out and processing terminates.[6] If no abortive condition is found, the BUILD program creates four output files: (1) the raw data file (RDXXXX.), (2) the keyword-author file (KAXXXX.), (3) the keyword-dictionary file (KDXXXX.), and (4) the author-dictionary file (ADXXXX.). Each of these four output files can be either a disk file or a magnetic tape file, depending upon the computer system hardware available and the costs of different storage modes. At present, a good configuration appears to be having RDXXXX. and KAXXXX. created as disk files and KDXXXX. and ADXXXX. created as magnetic tape files. The first two data files will be needed almost everytime the CARL System is used but the other two files are needed only when a dictionary (either keyword or author) is required or when certain types of corrections are being made using the CHANGE program. Finally, separate printed dictionaries are produced for keywords and authors.

As shown in Figure 6, the query-retrieval program (QUERY) uses the keyword-author file (KAXXXX.) to respond to demand terminal queries from a user. The desired keyword(s) is (are) typed on a terminal by the user. The QUERY program examines the KAXXXX. data file to identify all references with the appropriate keyword(s). A message indicating the number of references located is sent to the user on the terminal. The user is given the option of having the actual reference numbers listed or the entire text of the reference citation(s) listed. If the user elects to have the entire reference citation(s) listed, the RDXXXX. data file is required. If, on the other hand, the number of references initially indicated is too many, the user can narrow the scope of the search by specifying additional keywords.

The maintenance program (CHANGE) requires access to all four data files (RDXXXX., KAXXXX., KDXXXX., and ADXXXX.) as shown in Figure 7. If the change desired involves one or more corrections to the existing system, the appropriate data files are updated. If the maintenance activity involves adding new references to the system, new raw data are input (from cards or disk), certain editing checks are performed and all four data files are updated. Finally, the operator has the option of obtaining a post-change listing of all references changed (or added).

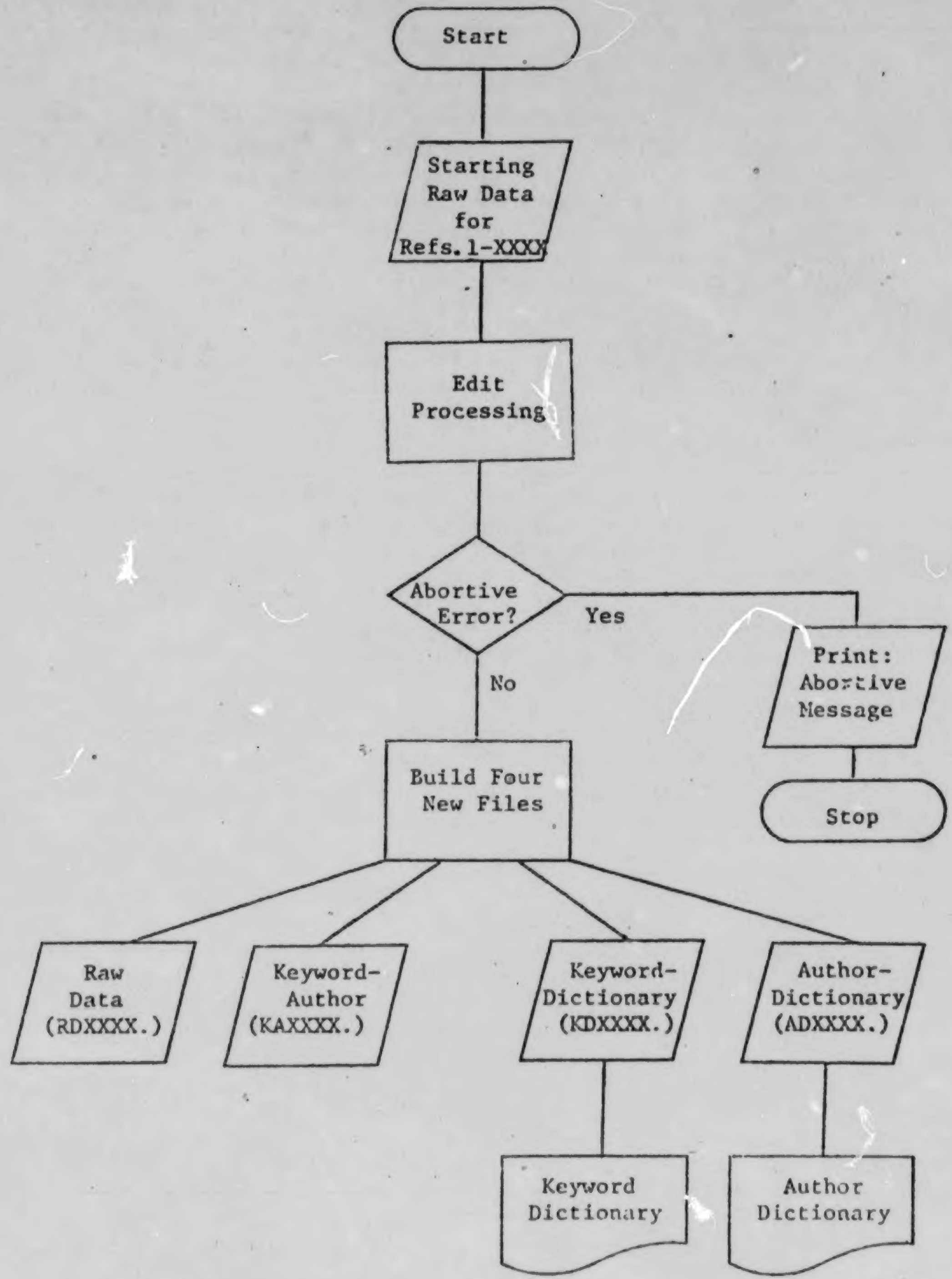
Query-Retrieval Example

The following example is presented to illustrate the way in which a user would interact with the CARL System. After contacting the person managing the system, the user would examine a controlled dictionary of all allowable keywords. This would enable the user to formulate the retrieval request in a manner which will be meaningful to the system.

---

[6] The operational definition of an abortive error condition can be specified by the system manager.

Program BUILD


Start
Starting Raw Data for Refs. 1-XXXX
Edit Processing
Abortive Error?
Yes
No
Print: Abortive Message
Stop
Build Four New Files
Raw Data (RDXXXX.)
Keyword-Author (KAXXXX.)
Keyword-Dictionary (KDXXXX.)
Author-Dictionary (ADXXXX.)
Keyword Dictionary
Author Dictionary

535

Figure 5. Program BUILD Flowchart.

Program QUERY

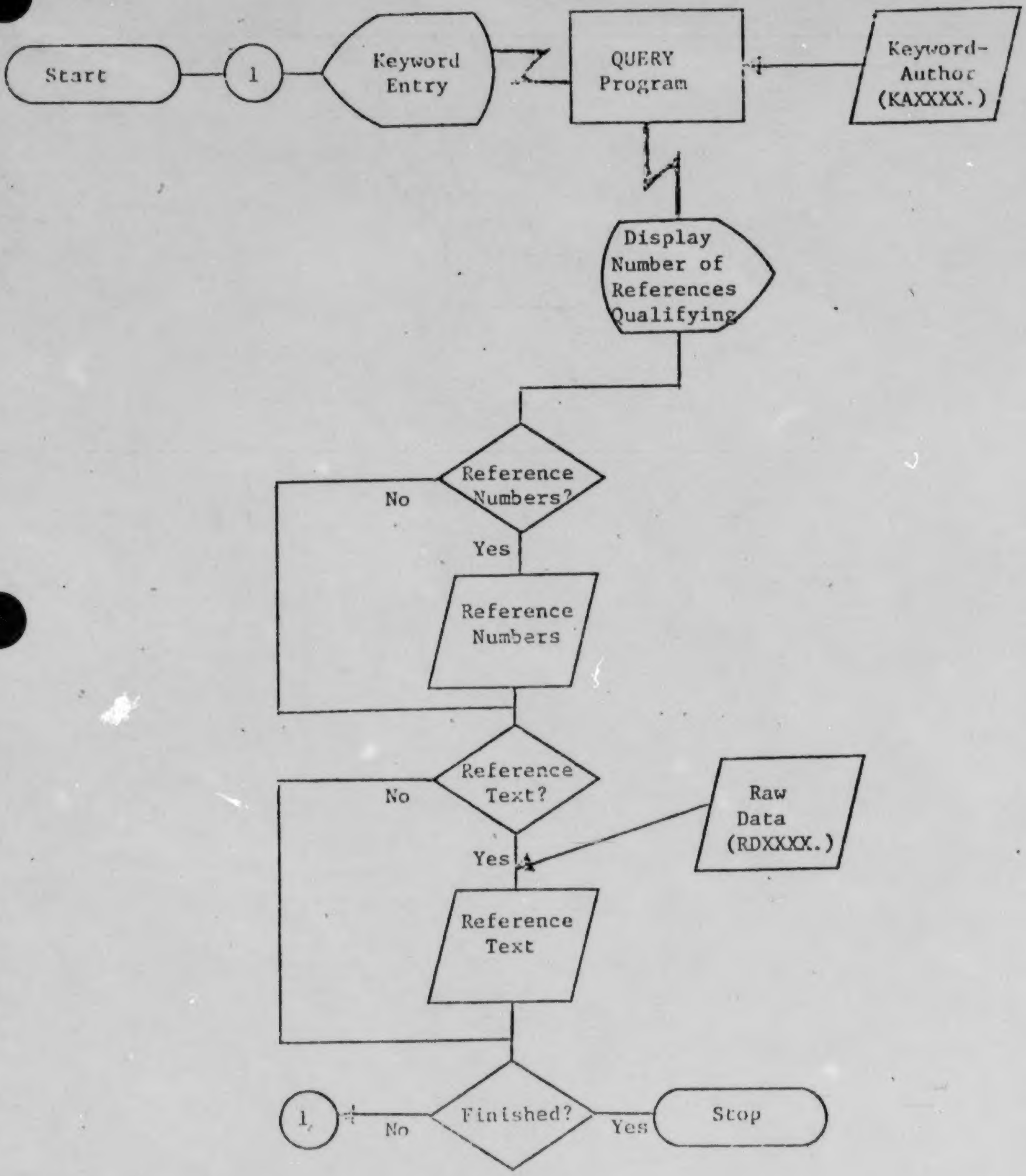

Start
1
Keyword Entry
QUERY Program
Keyword-Author (KAXXXX.)
Display Number of References Qualifying
Reference Numbers?
No
Yes
Reference Numbers
Reference Text?
No
Yes
Raw Data (RDXXXX.)
Reference Text
1
No
Finished?
Yes
Stop

Figure 6. Program QUERY Flowchart.

Program CHANGE

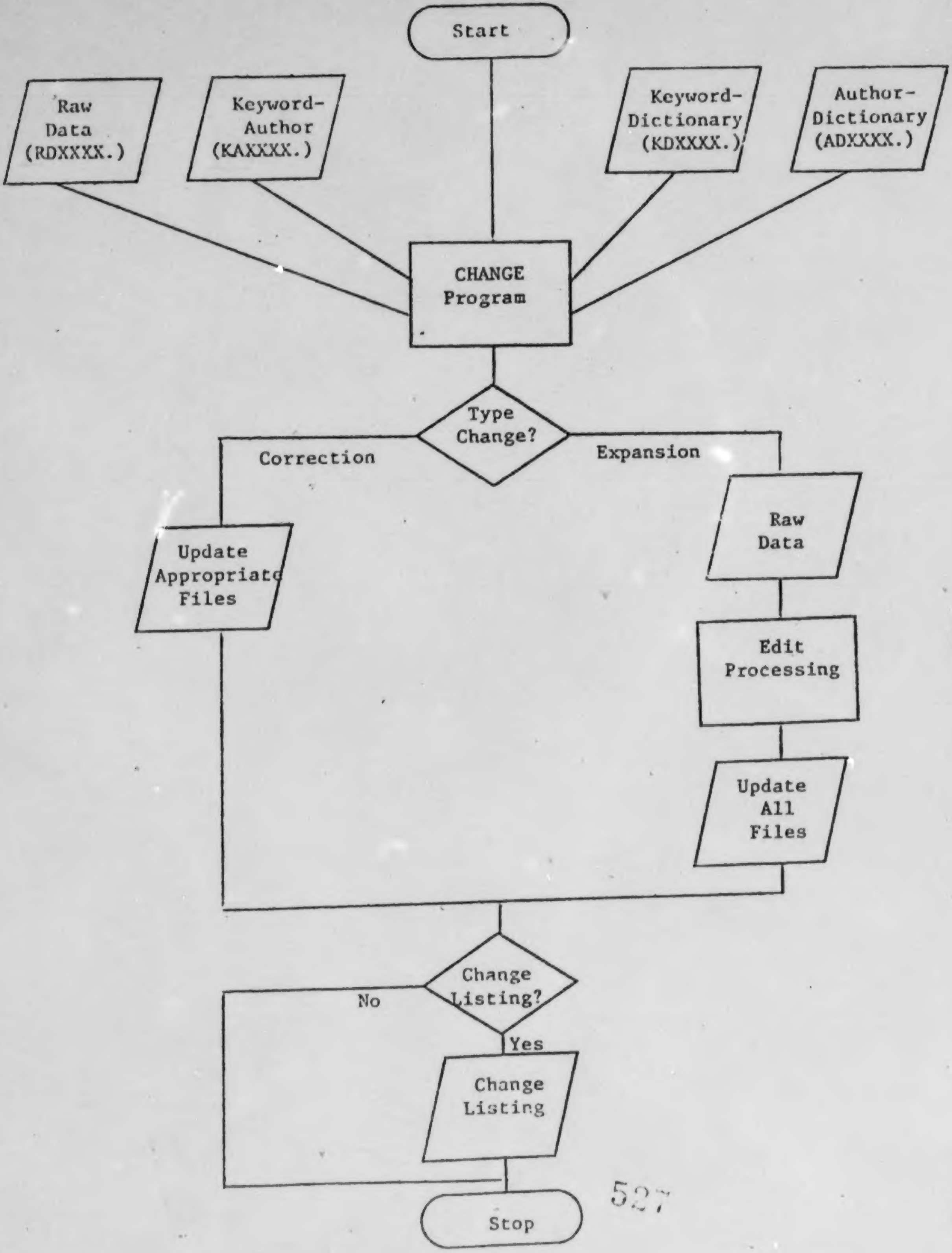

Start
Raw Data (RDXXXX.)
Keyword-Author (KAXXXX.)
Keyword-Dictionary (KDXXXX.)
Author-Dictionary (ADXXXX.)
CHANGE Program
Type Change?
Correction
Expansion
Update Appropriate Files
Raw Data
Edit Processing
Update All Files
Change Listing?
No
Yes
Change Listing
Stop

Figure 7. . Program CHANGE Flowchart.

Armed with the legitimate keywords which are pertinent to the topic, the user is ready to interact with the QUERY program of the CARL System, illustrated in Figure 6. After accomplishing the log-on procedures required to establish contact with the computer system, the user would enter the keyword(s) on a demand terminal. The QUERY program will locate all references which have been indexed with that keyword, or combination of keywords, and the number of references located will be displayed on the terminal. The user is then asked if a list of reference numbers is desired and, if so, the reference numbers are printed out in ascending sequence. Next the user is provided with the option of having the entire reference citation printed out for each reference located by the search. If the user wishes the entire reference citations, a hardcopy list will be produced in alphabetical sequence by author. Finally, the user is presented with the option of continuing the interaction with the CARL System or terminating the session and logging-off the computer system.

CONCLUSIONS

Coordinate Indexing

The coordinate indexing approach to information retrieval allows considerable flexibility. The source documents contained in the personal library system are not limited to published material available to the public as is the case with commercial reference retrieval systems. Lecture notes for teaching, documentation for computer programs, printed advertisements, equipment brochures, and notes used to present briefings are examples of the diversity which can be incorporated into a personal reference system. In addition, the coordinate indexing approach has considerable generality. For example, in the work setting, coordinate indexing could be employed in a management information system (MIS) designed to provide military laboratory managers with current information on all on-going research and development projects. The source document for this MIS could be DD-1498 Forms. In a home setting, coordinate indexing could be used to create, query-retrieve, and maintain files on a minicomputer. These home files might contain a photography collection of prints or slides or a record collection of albums or tapes. Unquestionably there are many examples, both in the office and the home, where an information retrieval system employing coordinate indexing could be quite useful.

Future Work

Eventually a disk-oriented version of the CARL System will be created. This second version of the CARL System will use direct access methods as opposed to the sequential access methods used by the original CARL System described herein. The direct access version will require considerably more programming time to develop than the sequential access version. However, once developed and debugged, the direct access version should significantly speed up the information retrieval process while simultaneously providing a substantial reduction in computer costs. Further, the advantage of a direct access system over a sequential access system will increase as the number of references in the system increases.

The possibility of a CARL System Network will be given consideration. This network would allow individual researchers to have access to the personal reference libraries of other researchers, in a mutually agreed upon fashion, thereby increasing the number of references examined for any query.

REFERENCES

Bourne, C. P. *Methods of Information Handling*. New York: John Wiley & Sons, Inc., 1963.

Broadhurst, P. L. Coordinate Indexing: A Bibliographic Aid. *American Psychologist*, 1962, *17*, 137-142.

Daniel, R. S., & Louttit, C. M. *Professional Problems in Psychology*. New York: Prentice Hall, Inc., 1953.

de Alarcón, R. A Personal Medical Reference Index. *The Lancet*, 1969, *1*, 301-305.

Lancaster, F. W. *Information Retrieval Systems: Characertistics, Testing, and Evaluation*. New York: John Wiley & Sons, Inc., 1968.

Price, D. J. S. *Science Since Babylon*. New Haven: Yale University Press, 1961.

Price, D. J. S. *Little Science, Big Science*. New York: Columbia University Press, 1963.

Redican, W. K. Reprints: File Before They Defile You. *American Psychologist*, 1973, *28*, 625-627.

Taube, M. *Studies in Coordinate Indexing*, Vol. 1. Washington, D.C.: Documentation Inc., 1953.

SECTION 5

PERSONNEL APPRAISAL

QUALITY OF ROTC ACCESSIONS TO THE ARMY OFFICER CORPS

by

Arthur C. F. Gilbert, Ph.D.
U. S. Army Research Institute for the Behavioral
and Social Sciences

John I. Weldon, Jr., Ph.D.
U. S. Army Training and Doctrine Command

Richard S. Wellins, Ph.D.
U. S. Army Research Institute for the
Behavioral and Social Sciences

A Paper Prepared for Presentation at the
20th Annual Conference of the Military Testing Association (MTA)
Oklahoma City, Oklahoma
October 30 - November 3, 1978

Performance and Training Research Laboratory
U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

QUALITY OF ROTC ACCESSIONS TO THE ARMY OFFICER CORPS

Arthur C. F. Gilbert, Ph.D.
U. S. Army Research Institute for the Behavioral and Social Sciences [1]

John I. Weldon, Jr., Ph.D.
U. S. Army Training and Doctrine Command

Richard S. Wellins, Ph.D.
U. S. Army Research Institute for the Behavioral and Social Sciences

The Army Reserve Officers' Training Program (ROTC) and the other officer procurement programs have to produce a sufficient number of officers to meet the requirements of the Army active and reserve components. Projections of future requirements appear to indicate that the ROTC program will need to double its number of graduates within the next few years in order to remain responsive to this need. As a consequence, the Professors of Military Science have been striving diligently to meet this objective by enrolling increasing numbers of students in the ROTC program. As in any personnel selection this, in turn necessitates a system, it is necessary to have a more stringent evaluation of the quality of accessions as more emphasis is being placed on quantity.

The objective of this research was to evaluate the quality of ROTC graduates and to determine if there were differences among ROTC graduates in performance on the basis of sex or on the basis of the geographical region in which the ROTC units are located. The criterion was the final course grades in the Officer Basic Courses (OBC) of the 13 Career Branches.

Procedure

A sample of 1,243 officers who completed Officer Basic Course in the first and second classes after 15 June 1977 were used in this research. In addition, a sample of 4,662 officers who continued on active duty after completion of OBC in Fiscal Year 1974 were selected from a total of 9,180 officers who entered on active duty during that year.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

Each sample was divided on the basis of sources of commission, ROTC, USMA, OCS, and direct appointments. The ROTC samples were divided into those officers who were ROTC scholarship recipients and those who were not. The ROTC sample for 1977 was divided into geographical regions corresponding to the location of the ROTC institution that they attended. Finally, the 1977 ROTC graduates were divided on the basis of sex.

Results and Discussion

The means of the four groups of officers from the four procurement programs are shown in Table 1 for the 1977 sample. Also, the means of the different subgroups are presented (i.e., ROTC Scholarship recipients, ROTC region and male and female samples.).

The results of the analysis of variance among the four procurement programs was significant. The average final OBC grades for the four procurement programs ranked as follows: U. S. Military Academy, ROTC, OCS, and direct appointments. Even though the U. S. Military Academy graduates were favored, a meaningful difference in mean performance for that group and for ROTC graduates was not obtained. When the ROTC graduates are classified as ROTC scholarship recipients and non-recipients, the mean Officer Basic Course final grades of the graduates of the different programs ranked as follows: U. S. Military Academy graduates, ROTC scholarship recipients, non-recipients of ROTC scholarships, OCS graduates, and direct appointments.

When an analysis of variance was performed to detect differences among the four ROTC regions, a significant difference was obtained. The Western Region was favored in terms of average OBC final grades earned while the South Central Region had the lowest average performance. A significant difference did not exist between the mean performance of male ROTC graduates on the criterion measure.

ROTC units who had five or more graduates in the 1977 sample were ranked on the basis of average OBC final grades. Of the 70 ROTC units ranked, the average OBC final grades of 18 of the 70 institutions so ranked exceeded that of the average OBC final grades of graduates of the U. S. Military Academy. The average OBC performance graduating of 50 of the ROTC institutions exceeded that of the average performance of OCS graduates while the average performance of graduates of 54 ROTC institutions exceeded that of the average performance of those officers who received direct appointments.

The means of the Officer Basic Course final course grades for each of the four procurement programs in the Fiscal Year 1973 sample are shown in Table 2 as well as the mean performance of ROTC scholarship recipients and non-recipients. Results of analysis of variance revealed a significant difference in performance among the four groups.

533

TABLE 1

MEANS OFFICER BASIC COURSE FINAL GRADES FOR THE DIFFERENT GROUPS IN THE 1977 SAMPLE OF OFFICER ACCESSIONS

Group	N	$\bar{X}$
U. S. Military Academy	113	106.90
ROTC	871	100.34
OCS	132	96.22
Direct Appointment	61	94.37
Total	1,177*	100.20
ROTC Scholarship Recipients	347	105.81
Non-Recipients	524	96.72
Male ROTC Graduates	814	100.48
Female ROTC Graduates	57	98.30
Eastern ROTC Region	341	98.13
North Central ROTC Region	188	101.99
South Central ROTC Region	155	96.83
Western ROTC Region	170	106.14

*All 1,243 cases were not used due to missing data elements.

TABLE 2

MEANS OF THE DIFFERENT PROCUREMENT PROGRAM GROUPS IN THE FY 1974 SAMPLE

Group	N	$\bar{X}$
U. S. Military Academy	591	99.04
ROTC	1,721	100.80
OCS	113	106.10
Direct Appointment	76	95.71
Total	2,501	100.47
ROTC Scholarship Recipients	598	102.54
Non-recipients	1,123	99.87

535

Graduates of OCS were favored over U. S. Military graduates and ROTC graduates while those officers who received direct appointments had the lowest mean OBC final course grades.

Those ROTC units who had five or more graduates in the Fiscal Year 1975 sample were rank-ordered on the mean Officer Basic Course final grades of the graduates. Inspection of these means revealed the means performance of 106 of the 235 institutions so ranked exceeded that of the mean performance of U. S. Military Academy graduates. The average performance of graduates of 37 ROTC institutions exceeded the average performance of OCS graduates in this sample while the average performance of 105 ROTC institutions exceeded the average OBC performance of officers who received direct appointments.

For the ROTC institution and that had five or more graduates both in the 1977 sample and in the Fiscal Year 1974 sample, the mean performance of the graduates were ranked within each sample. A Spearman rank order correlation coefficient was computed between the obtained values. The resulting correlation coefficient of .53 between these rankings was significant at the .01 level.

The results of this research indicate that the ROTC program is producing a quality of graduates whose performance in the Officer Basic Course is of comparable quality with other officer procurement programs. There appears to be a variability among the ROTC institutions in terms of the performance of graduates in Officer Basic Course but even so, the ROTC is meeting its objective of obtaining quality accessions for the officer corps. There appears to be a certain tendency for ROTC institutions that have produced graduates in Fiscal Year 1974 who performed well in Officer Basic Courses to do so again in 1977.

Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB)

John J. Mathews and Lonnie D. Valentine, Jr.
Personnel Research Division
Brooks Air Force Base, Texas 78235

Wayne S. Sellman, Major, USAF
Air Force Manpower and Personnel Center
(Research and Measurement Division)
Randolph Air Force Base, Texas 78148

Background

The General Accounting Office (GAO) submitted a report dated 31 March 1977 to the Secretary of Defense entitled ''A Need to Address Illiteracy Problems in the Military Services.'' Among other things, it recommended that the Department of Defense develop a policy to address the illiteracy problem and have the Services (1) determine the reading grade level required for each military occupation, and (2) establish an overall minimum reading level required for enlistment.

In a 10 June 1977 letter to the GAO, the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) concurred in general with the findings of the report (i.e., illiterate service personnel do have higher discharge rates, do experience more difficulty in training, and do have less potential for career advancement) but indicated that DOD's mission did not include the societal responsibility for remedying any deficiencies in the American educational system. Subsequent to the 10 June 1977 letter, other initiatives surfaced which were directly related to the illiteracy problem. The House and Senate Defense Appropriations Committees expressed concern about in-service high school completion programs and the potential impact of continuing to attempt to correct educational deficiencies of enlistees after they enter the Service. The Committees believed instead that a more efficient approach would be for potential enlistees with educational weaknesses to receive basic skills training prior to enlistment. Accordingly, the Secretaries of Health, Education, and Welfare (HEW) and Labor, in coordination with the Secretary of Defense, were requested to develop such a basic skills program.

Introduction

The result of these initiatives was increased OSD emphasis on the Services' literacy programs. In that regard the Principal Deputy Assistant Secretary of Defense (Manpower, Reserve Affairs, and

537

Logistics) directed by memorandum, dated 18 October 1977, that a ''study be conducted to evaluate the capability of the Armed Services Vocational Aptitude Battery (ASVAB) to determine the reading ability skills of applicants for enlistment at the Armed Forces Examining and Entrance Stations (AFEESs).'' It was believed that because of its highly verbal content, the ASVAB already indirectly measured reading ability. If that was, in fact, the case, most applicants with low reading skills were already being screened out. In addition, if a reading grade index could be derived from ASVAB, estimates of applicants' reading skills could be provided to Labor and HEW representatives involved in the programs alluded to above.

Thus, the specific objectives of this study were to assess the reading ability of applicants for military service as well as for actual accessions and to determine the relationship between ASVAB measures (Jensen, Massey, & Valentine, 1976) and reading scores. Depending on the magnitude of the relationship, an appropriate combination of ASVAB subtests could be used to estimate the reading grade level of groups of applicants and possibly to predict within a reasonable confidence interval the reading grade level of individuals. The present report concerns analyses involving two reading tests. Additional data covering two other reading tests will be presented in a subsequent report.

Method

Subjects

The study plan called for testing 6,000 service applicants divided among 25 geographically dispersed AFEESs. Four reading tests were administered, the Gates-MacGinitie, Nelson-Denny, Basic Skills Assessment, and Literacy Assessment Battery, with each subject taking two of the tests. This report concerns all subjects given the Gates-MacGinitie tests and a subsample who were also given the Nelson-Denny test. In March-April 1978, 2,899 applicants were given the Gates-MacGinitie test, and ASVAB scores obtained for 2,432 of these. The first sample consists of 2,033 of the 2,432 for whom sufficient identification was available from reading and ASVAB data sources to obtain accurate matches, and for whom most other data of interest (e.g., sex, race, education) was also valid. A subsample consists of 818 of the 2,033 who were given the Nelson-Denny reading test in addition to the Gates-MacGinitie. The second sample includes 212 subjects who took the Gates-MacGinitie and Nelson-Denny, but for whom no ASVAB data were available. Reading data for these was compared to that for the 818 to detect possible bias in the samples.

Predictors

An Applicant Processing Worksheet was available for most of the subjects. ASVAB subtest scores and Armed Forces Qualification Test (AFQT) percentiles were obtained from these documents. Other analysis variables from the worksheets included military service applied for, educational level, race, sex, and service qualification status--qualification being a function of an applicant's meeting specified minimum ASVAB and educational criteria. Sample percentages for demographic variables are in Appendix A.

Criteria

The reading tests involved in this report were the Gates-MacGinitie Reading Tests Survey D (Gates & MacGinitie, 1965) and the Nelson-Denny Reading Test Form C (Brown, 1973). The order of administration of these tests was counterbalanced. Both tests contain a vocabulary and a reading comprehension subtest which were separately scored. The published test norms were used to convert the reading test raw scores to reading grade level scores.

Statistical Methods

Statistical analyses included multi-variate distributions and correlation matrices. Due to a difference in range and distributions, reading grade levels for the two reading tests have been summarized in most instances by use of medians rather than means. The best combinations of ASVAB subtests for predicting reading levels was determined via multiple regressions.

Results and Discussion

Percentages of service applicants scoring at each reading grade level as measured by the Gates-MacGinitie test are shown on the right side of Table 1. The reading grade level range of Gates-MacGinitie which is targeted at 4th-6th grades is from 2 to 11. The top reading grade level, labeled ''11 & above,'' contains the largest proportion of applicants, 565 or 27.8% of 2,033. About 7.8% obtained reading grade levels below four. The median reading grade level of applicants was 9.0.

Due primarily to aptitudinal and educational screening standards employed by services, the reading grade levels of examinees meeting the qualification standards of the service for which tested were usually higher than those of examinees who did not qualify. The median reading grade level of applicants qualifying for services was 10.2 compared to 5.7 for non-qualifying applicants.

Since each service has different screening standards and uses different combinations of abilities, the aptitude and education distributions vary across services for applicants and especially for accessions. This is reflected in relatively higher reading grade levels for Air Force and Navy applicants than for Army and Marine Corps applicants. As indicated in Table 1, the median reading grade level for applicants qualifying for the Air Force was 10.9 and the median reading grade level for those qualifying for the Navy was 10.5, while the median reading grade level for Army and Marine Corps qualified applicants was 9.3 each.

The impact of completion of high school on reading grade level can be seen in Table 2 which gives percentages of graduates and non-graduates at each reading grade level. The median reading grade level for high school graduates was 9.8 compared to 7.9 for high school non-graduates. The effect of aptitude screening on reading grade level is also evident from data in Table 2. High school graduates who qualified for services had a median reading grade level of 10.6 while high school graduates who did not qualify had a median reading grade level of 6.1.

The Armed Forces Qualification Test (AFQT) which is used for preliminary screening by all services was correlated with the Gates-MacGinitie. The correlation (r) between AFQT percentiles and reading grade level was .74. For the Black applicants in the sample (N = 835) the r was .68 (race and sex distributions of reading grade level appear in Appendix B). To gauge the magnitude of this relationship, the construct validity and reliability of the Gates-MacGinitie and the reliability of AFQT must be considered. Due to less than perfect reliability of these measures, their maximum intercorrelation would be less than one.

Data for a subsample of the 2,033 who had also taken the Nelson-Denny reading test (N = 818) was analyzed for additional information. The 818 appeared to be representative of the 2,033, with mean Gates-MacGinitie reading grade levels of 8.6 and 8.4, respectively, and a common Standard Deviation of 2.8.

The Nelson-Denny has a reading grade level range of from 6 to 15 and is targeted at about the 11th-13th grades. Table 3 contains comparable data for samples for which Gates-MacGinitie and Nelson-Denny data were analyzed. The median reading grade level for Nelson-Denny was 9.5 compared to 9.0 for Gates-MacGinitie. While 32.4% of applicants had Gates-MacGinitie reading grade levels of six or less, only 10.8% of applicants had Nelson-Denny reading grade levels of six or less. The mean AFQT percentile of those with reading grade levels of six or less was 25.5 for Gates-MacGinitie and 31.9 for Nelson-Denny. The correlation between Nelson-Denny reading grade level and AFQT was .65 compared to the r of .74 between Gates-MacGinitie and AFQT (intercorrelations of reading tests, AFQT, and selected ASVAB subtests are listed in Table 4.)

The r between the average of Gates-MacGinitie and Nelson-Denny reading grade levels and AFQT was .76.

The intercorrelation between Gates-MacGinitie and Nelson-Denny reading grade levels was .69. If these tests are measuring the same ability (reading), then AFQT is also measuring reading with comparable precision since AFQT correlates to about the same degree with Gates-MacGinitie and Nelson-Denny as these reading tests do with each other.

AFQT is not the best ASVAB measure of either reading grade level, however. Not surprisingly, the ASVAB subtest with the highest relationship to reading scores was Word Knowledge (WK). This vocabulary test correlated .73, .69, and .78 with Gates-MacGinitie, Nelson-Denny, and the average of the two reading grade levels, respectively. Of the other two subtests (besides WK) which form the AFQT, Arithmetic Reasoning (AR) correlated substantially higher with reading grade level than did Space Perception (SP). The r between AR and average reading grade level was .62, compared to .35 between SP and average reading grade level. This indicates that a composite of WK and AR (the General Technical composite used by Army and Navy, and the General composite used by Air Force) would be an even more valid predictor of reading grade level than AFQT. The General Technical composite (GT) correlated .76, .68, and .79 with Gates-MacGinitie, Nelson-Denny, and average reading grade levels, respectively. Compared to the r of .76 between AFQT and average reading grade level, GT accounts for about 8% more variance in reading grade levels than does AFQT.

Based on multiple correlation (R's), the best two ASVAB subtest combination for predicting both reading tests consisted of WK and Numeric Operations (NO), a clerical speeded subtest. The R's of WK and NO were .77, .75, and .83 with Gates-MacGinitie, Nelson-Denny, and average reading grade levels, respectively. The three ASVAB subtest combination which correlated highest with reading grade levels included General Science (GS). The R's of WK, NO, and GS with Gates-MacGinitie, Nelson-Denny, and average reading grade level were .80, .77, and .86.

The choice among commercial reading tests and some combination of ASVAB measures as optimal for estimating reading grade levels of service applicants should be based on considerations involving fairness, difficulty levels, and administrative considerations as well as validity and reliability. The reading tests (Gates-MacGinitie + Nelson-Denny) correlated slightly higher with race than did AFQT (-.44 vs. -.37). Minorities did relatively less well on both reading tests than on AFQT. Gates-MacGinitie plus Nelson-Denny also had a higher r with the dichotomous variable sex than did AFQT (.19 vs. .10). Females scored higher on both AFQT and reading tests, but this sex difference was less on AFQT.

Regarding difficulty levels, the form of Gates-MacGinitie used would be appropriate for minimum cutoff scores around 4th-6th reading grade levels. However, Gates-MacGinitie would be too easy for cutoffs at the 9th reading grade level (used by the Air Force) or for accurate estimates of group reading grade levels since the median of service accessions was only one grade lower than the top Gates-MacGinitie reading grade level. The Nelson-Denny form used would be too difficult for use for cutoffs around the 4th-6th reading grade levels since the sixth grade was the lowest Nelson-Denny reading grade level. The ASVAB was developed for the service applicant population. The mean item difficulty level (proportion of examinees correctly answering items) is about .6 on AFQT and GT (uncorrected for guessing).

From an administrative standpoint, the easiest way to obtain estimates of reading grade level would be currently used ASVAB composites (AFQT or GT). An unweighted combination of ASVAB subtests (such as WK + GS + NO) would be somewhat less convenient and probably not much more valid. A weighted composite of WK + GS + NO would give a somewhat better estimate of reading grade level, but would require additional computations. A reading grade level index computed from ASVAB could be used to tailor basic skills remediation programs to the reading levels of their referrals.

The sample of 818 taking the Gates-MacGinitie and Nelson-Denny tests was compared to 212 who also took these tests but for whom no ASVAB data were available. It had been speculated that many of those without ASVAB data were of marginal aptitude and did not return to take the ASVAB after doing poorly on the reading tests. This was not the case, however, as the mean average reading grade level was slightly higher for the 212 than for the 818 (9.8 vs. 9.4).

Conclusions

The main findings of this study were:

1. The median reading grade level for service applicants was 9.0 based on Gates-MacGinitie and 9.5 based on Nelson-Denny. The median Gates-MacGinitie reading grade level of applicants who qualified for services was 10.2 compared to 5.7 for non-qualified applicants.

2. The AFQT correlated .74 with Gates-MacGinitie, .65 with Nelson-Denny, and .76 with average reading grade levels, respectively. Since the intercorrelation of Gates-MacGinitie and Nelson-Denny was .69, AFQT appeared to measure reading as well as the reading tests. The GT composite (General AI for Air Force) correlated .79 with average reading grade level.

3. The multiple correlations between the three ASVAB subtest combination of WK, GS, and NO, and the Gates-MacGinitie, Nelson-Denny, and average reading grade levels were .80, .77, and .86, respectively.

4. ASVAB is presently screening out most applicants with marginal literacy skills.

Recommendations

The GT composite of ASVAB should be used as an index of reading grade level. A conversion table can be developed for predicting reading grade levels from GT scores.

REFERENCES

Brown, J.I., Nelson, M.J., & Denny, E.C. The Nelson-Denny Reading Test. Examiner's Manual. Boston MA: Houghton Mifflin Company, 1976.

Gates, A.I., & MacGinitie, W.H. Gates-MacGinitie Reading Tests, Survey D. Teacher's Manual. New York: Teachers College Press, 1965.

Jensen, H.E., Massey, I.H., & Valentine, L.D., Jr. Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7). AFHRL-TR-76-87, AD-A037 522. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Table 1. Percentages of Qualified and Not Qualified Applicants by Service at Each Gates-MacGinitie Reading Grade Level

Reading Grade Level	Qualified					Not Qualified					All Applicants		
	Army	Navy	AF	MC	All	Army	Navy	AF	MC	All	RGL	%	N
11 & above	30.7	43.1	48.9	24.8	37.8	0.7	5.6	5.2	-	2.4		27.8	565
10-10.9	11.9	14.9	19.2	13.8	14.3	5.2	5.6	12.9	3.5	7.1		12.3	249
9-9.9	10.2	10.0	12.9	15.2	11.2	3.1	7.0	9.7	3.5	5.4		9.5	194
8-8.9	8.6	10.1	6.9	11.0	8.9	6.6	9.9	7.7	6.9	7.3		8.5	172
7-7.9	9.8	9.4	6.3	13.1	9.3	7.6	11.3	14.8	10.3	10.3		9.5	194
6-6.9	11.1	5.5	1.9	9.7	7.3	12.4	16.9	13.6	8.6	12.9		8.9	180
5-5.9	9.8	3.7	1.9	6.9	6.0	21.0	12.7	14.2	17.2	17.8		9.3	189
4-4.9	5.2	1.8	0.6	1.4	2.8	15.9	16.9	12.6	22.4	15.7		6.4	131
3-3.9	2.1	1.4	0.6	2.8	1.6	13.8	9.9	6.5	8.6	10.8		4.2	86
2.9 & below	1.6	0.2	0.6	1.4	1.0	13.8	4.2	3.2	19.0	10.3		3.6	73
Total Percent	100	100	100	100	100	100	100	100	100	100		100	
Median Reading Grade Level	9.3	10.5	10.9	9.3	10.2	5.3	6.4	7.0	5.0	5.7	9.0		
Total N	561	436	317	145	1,459	290	71	155	58	574			2,033

Table 2

Percentage of High School Graduates and non-Graduates at Each Gates-MacGinitie Reading Grade Level by Qualified/non-Qualified

Estimated Reading Grade Level	High School Graduate			High School non-Graduate		
	Qualified	Not Qualified	All Grad	Qualified	Not Qualified	All Non-Grad
11 & above	42.9	3.7	34.3	30.3	1.5	20.0
10-10.9	15.8	6.7	13.8	12.0	7.6	10.4
9-9.9	11.0	5.4	9.8	11.6	5.5	9.4
8-8.9	7.6	8.7	7.8	10.8	6.1	9.1
7-7.9	8.1	12.4	9.0	11.1	8.8	10.3
6-6.9	4.7	15.3	7.0	10.8	11.0	10.9
5-5.9	5.0	14.9	7.2	7.0	19.8	11.6
4-4.9	2.5	16.1	5.5	3.4	15.5	7.8
3-3.9	1.6	9.9	3.5	1.7	11.6	5.3
2.9 % below	0.8	7.0	2.2	1.2	12.5	5.3
Total Percent	100	100	100	100	100	100
Median Reading Grade Level	10.6	6.1	9.8	9.3	5.5	7.9
Total N	855	242	1,097	584	328	912

Table 3. Comparison of Reading Grade Level and AFQT for Gates-MacGinitie (N = 2,033) and Nelson-Denny (N = 818) Samples

Reading Grade Level	Cumulative %		AFQT Mean			
	Gates-MacGinitie	Nelson-Denny	Gates-MacGinitie		Nelson-Denny	
15 & above	-	100	-		81.9	
14-14.9	-	94.8	-		76.0	
13-13.9	-	88.1	-		64.5	(66.9)[2]
12-12.9	-	78.7	-		57.5	
11-11.9	100	70.1	70.9		60.8	
10-10.9	72.2	63.9	55.1		49.6	
9-9.9	59.9	55.0	50.9		46.9	
8-8.9	50.4	42.4	46.4		40.4	
7-7.9	41.9	27.7	38.8		38.2	
6-6.9	32.4	10.8	32.0		31.9	
5-5.9	23.5	-	28.9		-	
4-4.9	14.2	-	22.7	(25.5)[1]	-	
3-3.9	7.8	-	18.3		-	
2.9 & below	3.6	-	14.2		-	
Median Reading Grade Level	9.0	9.5				
AFQT Mean			47.2		50.1	
Standard Deviation			23.7		22.5	
Total N	2,033	818				

---

[1] Mean for 6 and below

[2] Mean for 11 and above

546

Table 4

Means, Standard Deviations, and Intercorrelations of Variables for Gates-MacGinitie + Nelson-Denny Subsample (N = 818)

	Mean	SD	Intercorrelations 1	2	3	4	5	6	7	8	9	10	11	12	13
1 Sex[1]	1.16	.37	1.00	-.01	.20	.10	.11	.19	.14	.02	-.11	.04	.15	.20	.19
2 Race[2]	1.43	.55	-.01	1.00	.02	-.37	-.38	-.30	-.34	-.35	-.22	-.37	-.40	-.40	-.44
3 Education Level	11.58	1.27	.20	.02	1.00	.28	.29	.25	.30	.24	.06	.31	.23	.36	.32
4 AFQT Percentile	50.06	22.52	.10	-.37	.28	1.00	.94	.57	.88	.82	.62	.73	.74	.65	.76
5 GT Percentile	54.33	27.20	.11	-.38	.29	.94	1.00	.58	.94	.83	.38	.68	.76	.69	.79
6 NO	30.56	10.19	.19	-.30	.25	.57	.58	1.00	.49	.58	.28	.49	.58	.59	.64
7 WK	18.63	6.71	.14	-.34	.30	.88	.94	.49	1.00	.62	.33	.71	.73	.69	.78
8 AR	11.47	4.30	.02	-.35	.24	.82	.83	.58	.62	1.00	.40	.58	.60	.54	.62
9 SP	12.06	3.92	-.11	-.22	.06	.62	.38	.28	.33	.40	1.00	.43	.39	.25	.35
10 GS	10.31	3.91	.04	-.37	.31	.73	.68	.49	.71	.58	.43	1.00	.70	.67	.74
11 Gates-MacGinitie Reading Grade Level	8.60	2.82	.15	-.40	.23	.74	.76	.58	.73	.60	.39	.70	1.00	.69	.92
12 Nelson-Denny Reading Grade Level	10.09	2.73	.20	-.40	.36	.65	.69	.59	.69	.54	.25	.67	.69	1.00	.92
13 Average Reading Grade Level[3]	9.37	2.55	.19	-.44	.32	.76	.79	.64	.78	.62	.35	.74	.92	.92	1.00

[1] Male = 1, Female = 2

[2] Caucasian - 1, Minority = 2

[3] Average of Gates-MacGinitie and Nelson-[illegible]y Reading Grade Levels for each subject.

APPENDIX A

Frequency Distributions of Variables for Gates-MacGinitie Sample (N = 2,033) and Nelson-Denny Subsample (N = 818)

	Gates-MacGinitie Sample		Gates-McGinitie + Nelson-Denny Subsample	
	N	%	N	%
Service				
Army	851	41.9	371	45.4
Navy	507	24.9	187	22.9
Air Force	472	23.2	195	23.8
Marine Corps	203	10.0	65	8.0
Race				
White	1,198	58.9	508	62.1
Black	835	41.1	310	37.9
Sex				
Male	1,652	81.3	688	84.1
Female	381	18.7	130	15.9
Qual. Status				
Qualified	1,459	71.8	645	78.9
Not Qualified	574	28.2	173	21.1
AFEES				
Atlanta	273	13.4	273	33.4
Boston	27	1.3		
Cincinnati	175	8.6		
Dallas	271	13.3	271	33.1
Fresno	89	4.4		
Indianapolis	196	9.6		
Jacksonville	35	1.7		
New Orleans	193	9.5		
Oklahoma City	189	9.3	189	23.1
Philadelphia	446	21.9		
Pittsburgh	85	4.2	85	10.4

APPENDIX B

Percentages of Applicants at Each Gates-MacGinitie Reading Grade Level by Race and Sex

Reading Grade Level	White	Black	Male	Female
11 & above	38.8	12.0	26.3	34.4
10-10.9	15.4	7.8	11.6	15.2
9-9.9	10.9	7.7	9.0	12.1
8-8.9	8.4	8.5	8.2	9.7
7-7.9	7.9	11.9	9.4	10.0
6-6.9	6.8	11.9	9.4	6.6
5-5.9	4.9	15.6	9.8	7.4
4-4.9	3.2	11.1	7.4	2.4
3-3.9	2.3	7.1	4.8	1.8
2.9 & below	1.5	6.6	4.3	0.5
Total percent	100	100	100	100
Median Reading Grade Level	10.3	6.8	8.6	10.0
Total N	1,198	835	1,652	381

550

SECTION 6

USING RATING SCALES

551

The Content Issue in Performance Appraisal Ratings

by

Randy H. Massey, Captain, USAF
C. J. Mullins
James A. Earles
Personnel Research Division
Brooks Air Force Base, Texas

Introduction

Much research done on ratings has been concerned with efforts to determine the best stimulus statements to use in a rating situation. Unfortunately, in much of this research ‘‘best’’ has been defined in terms of psychometric properties inherent in the ratings. Little research has been done employing external criteria for evaluating rating statements. This study focuses on the relative merits of rating statements with content selected to represent different points on a continuum from highly job-specific statements to person-oriented, trait-like statements. A context was constructed which provides an opportunity to evaluate the usefulness of various sets of rating statements against criteria external to the ratings, rather than the more traditional method of evaluating rating statements in terms of their internal psychometric characteristics.

The generally accepted viewpoint is that the more specific observable behaviors are more accurately rated than general personality descriptive statements. This viewpoint appears to be based more on the selective appraisal of a narrow spectrum of studies rather than on an appraisal of all studies conducted in the field (Kavanagh, 1971). In any case, the difficulties and controversial issues inherent in ratings have been well documented (e.g., Barrett, 1966; Kavanagh, 1971; Ronan & Prien, 1971; Schmidt & Kaplan, 1971).

Three prominent methodological procedures in developing rating stimulus statements or evaluation attributes include the following approaches: Behavioral Expectation Scales (Smith & Kendall, 1963); multitrait-multimethod (Campbell & Fiske, 1959); and McCormick's (1957) job analysis approach.

In the Behavioral Expectation Scales (BES) approach, important performance dimensions are identified and defined by a group of individuals responsible for evaluations. The scales are anchored by actual job behaviors which represent specific performance levels. The multitrait-multimethod approach uses data from many traits and raters which are analyzed for convergent and discriminant validity. The optimum stimulus statements should possess high convergent validity correlation coefficients and low discriminant validity correlation coefficients (Campbell & Fiske, 1959). McCormick (1957) emphasizes the importance of using job-oriented and worker-oriented statements derived from job analysis techniques. Job-oriented statements describe the job content, or what is accomplished by the worker (repair water pump, inspect lubrication system, drive pick-up truck, etc.). Worker-oriented statements tend to characterize generalized human behaviors or worker characteristics which are usually descriptive across many different jobs (observe visual displays, judge condition or quality, manually pour ingredients into container, etc.).

Perhaps the most popular scaling procedure designed to measure job performance is the BES methodology developed by Smith and Kendall (1963). BES has had considerable intuitive appeal, and there have been many proponents of the technique (e.g., Campbell, Dunnette, Arvey, Hellervik, 1973; Campbell, Dunnette, Lawler, & Weick, 1970; Dunnette, 1966; Landy, Farr, Saal, & Fretag, 1976; and Zedeck & Blood, 1974). BES scales have also been developed for a variety of occupations (e.g., Arvey & Hoyle, 1974; Landy, et al., 1976; Smith & Kendall, 1963). However, a review of studies in which BES was compared to other formats does not provide support for the effectiveness of the BES methodology (e.g., Buranaska & Hollmann, 1974; Dickenson & Tice, 1973; Zedeck & Baker, 1972; Borman & Vallon, 1974).

Intrinsic to the BES methodology is the assumption of the superiority of behavior-based attributes over trait-oriented attributes. McCormick's (1957) job analysis approach assumes the superiority of behavior-based attributes as well as task-oriented attributes. The multitrait-multimethod approach is the only methodology that does not implicitly assume the superiority of behavior-oriented attributes over trait-oriented attributes. In fact, both types of attributes have been found to be effective in performance evaluation devices (Kavanagh, 1971) when employing the multitrait-multimethod approach. Considering the popularity of behavior-oriented statements, it is not surprising that the common belief is that behavior-based rating statements are superior to trait-oriented statements. Nevertheless, there is no comparative evidence to indicate the superiority of any of the aforementioned methodologies.

A common issue underlying all rating methodological approaches is the ''content issue'' defined by Kavanagh (1971), as ''the issue of the relative representativeness of traits . . . along a continuum ranging from subjective to objective, abstract to concrete, or personality to performance.'' He concluded that there is no overwhelming evidence to indicate the superiority of behavior-based over trait-oriented dimensions. He further suggests that contradictory findings across reliability and validity studies could be partially attributed to a failure to resolve or control for the ''content issue.'' Resolution of this issue may give insight into the effectiveness of various performance evaluation methodologies, particularly in relation to time and cost expended. Settlement of this issue can also have significant explanatory value accounting for the numerous contradictory findings that exist in performance appraisal research.

Kavanagh, MacKinney, and Wollins (1971) were the first to directly address the content issue, using the multirater-multimethod approach, by investigating middle managers using performance ratings from superiors and two subordinates. They found more convergent validity for personal traits than performance traits, but no difference for discriminant validity. Although the higher personal trait convergent validity was accompanied by a greater degree of ‘‘halo,’’ the overall conclusion was that ratings of personal traits did as well as the ratings of performance traits.

Since Kavanagh (1971), the content issue has been almost entirely ignored. Recently Borman and Dunnette (1975) attempted to resolve the content issue by comparing behavior-based statements with trait-oriented statements. Their conclusions were, ‘‘at present little empirical evidence exists supporting the incremental validity of performance ratings made using behavior scales.’’ Unfortunately, there are methodological problems associated with their study. They compared three different rating systems (performance anchored, performance non-anchored, and trait-oriented statements obtained from the Naval Officer Fitness Report), rather than just comparing three rating formats. In sum, the study did not directly focus on the content issue of rating criteria, but rather on the effectiveness of three different rating systems. Among other experimental difficulties, they compared different numbers of rating statements between treatments and included trait-like statements (integrity, responsibility, and dedication) within the performance treatment category.

It seems clear, then, that the issue of the preferred content for rating statements has in no way been resolved by previous research. This study is one in a series of studies using criteria external to the ratings to attempt such a resolution. It is anticipated that this approach will be more effective in resolving the content issue than were past studies that employed internal characteristics of the rating instrument as criteria for judging the excellence of rating statements.

Method

Sample

One hundred and twenty students assigned to the ATC NCO Academy at Lackland AFB Annex completed the rating tasks. The study included nine separate seminar groups, each consisting of 13 or 14 NCOs (E6s to E7s) whose length of military service was 10 to 17 years.

Rating Scales

The treatment conditions in this study varied across three different types of rating statements (task-oriented, worker-oriented, and trait-oriented). Ten rating statements representing each of the three different kinds of rating content were included in the study. These were determined by consultation with instructors, administrative officials, and students. Previously conducted studies were also reviewed to identify factors. Each of the 10 rating attributes was rated on a 5-point scale as follows:

	Below Average	Average	Above Average	Well Above Average	Outstanding
Specific Ratable Attribute					

Trait oriented attributes also included a brief descriptive definition. See Appendix A for a complete list and description of the rating statements.

Rating Tasks

The research was conducted in two phases. In Phase I, each student rated all members in his seminar group on one and only one of the three different types of statements--task-oriented, worker-oriented, and trait-oriented. This phase resulted in the generation of individual profiles based on the group's evaluation of each member on each of the 10 selected rating attributes.

In Phase II, about 2 weeks later, the experimenter handed out the profiles to the seminar group without an identifying name on the profiles. Each subject was required to perform three tasks: first, he had to rank-order the profiles according to predicted seminar class rank; second, he had to identify to whom each profile belonged; and third, he had to predict the final school seminar class rank of his seminar peers without any regard to profile considerations. Subjects appeared unaware of the nature of the study until Phase II research when they were asked to identify each of the profiles.

Research Approach and Rationale

Many studies into the relative efficiency of sets of rating statements have apparently started with a basic set of assumptions: (1) Raters are subject to leniency error resulting in elevated means and to halo error revealed by small standard deviations among the ratings assigned. Since these two forms of rating error are revealed by the indicated statistics, a study of means and standard deviations forms a basis for comparison among sets of rating statements which may be used to distinguish among sets as to their goodness; (2) If rating statements are meaningful, and if raters are accurate in their perceptions of ratees, then inter-judge agreement, in the form of correlations among sets of ratings issuing from different judges, will be an expression of the goodness of a set of ratings; (3) The most useful way to compare sets of rating statements with each other lies in the comparisons which can be made among the summary statistics produced by the ratings. If one accepts these assumptions, then it follows that the best way to compare sets of rating descriptions is as it has frequently been done--the best set is that set which produces lower means, larger standard deviations, and larger inter-judge correlation coefficients.

However, the foregoing assumptions are subject to challenge. Taking them in order: (1) The evidence seems clear that leniency and halo errors do occur. It is less clear how important these two errors are in a family of other possible errors (e.g., racial bias, low rater motivation, low observability of the ratee, and others). It is also clear that there is not a direct relationship between leniency error and larger means or between halo error and smaller standard deviations. A person who is good on one dimension is more likely also to be good on whatever other dimensions are being considered. This is true whether the ''goodness'' metric is derived from ratings, from tests, or from any other reasonable source. Therefore, some portion of ''halo error'' may reflect true conditions, and be no error at all. (2) Inter-judge agreement may sometimes be a sufficient basis for comparing sets of rating statements, but it is not unusual for groups of judges to agree on a decision which additional facts show to be in error. If one may postulate individual differences among raters in respect to their ability to perceive ratees accurately, which seems plausible, then one must agree that some raters will provide better ratings. If some raters are better than others, it seems naive to expect that their ratings of a given characteristic will fall eternally at the mean of ratings given on that characteristic. (3) In this study, an approach is taken which provides a better basis for making comparisons across rating sets than does the traditional psychometric comparison. The approach is constructed around the concept of ''hits;'' that is, the number of times a rater can correctly identify anonymous profiles of his peers, constructed around various sets of descriptor statements.

If a rating statement is useful in describing a person, and if a group of raters can agree to some extent on the elevation of this characteristic in a ratee, then a profile of this ratee produced from a set of such statements should be identifiable as a rating ''picture'' of that individual. If a group of raters can recognize the individuals whom their profiles describe, then it seems more likely that the set of profiled characteristics can be useful in evaluating or predicting the performance of those individuals. The number of ''hits'' (correctly labeled profiles) should be useful in comparing one set of rating descriptions with another.

One analysis was made using hits as the dependent variable. The number of hits, however, at least in prior research (Curton, Ratliff, & Mullins, 1977), has proved so small that something more sensitive was needed. A rater could conceivably misidentify the first profile considered; and that misidentification could cause him to miss the rest, even if only by a small margin--or he could be so insensitive to personal differences that he makes guess errors in all the

556

identifications. The search for a sensitive measure of profile identification led to the use of the rank-order correlation as a possibly more effective measure of identification of peers than the simple count of correct identifications.

If a rater trying to identify anonymous profiles of his peers is confronted with 15 profiles, three of which have been rated very high on a particular characteristic, and if he believes correctly that peers B, H, and J are the three in his peer group highest on this characteristic, he may not know which of the three is peer B. He might specifically misidentify all three profiles, although he has been correct in believing that these three profiles, as a set, represent peers B, H, and J. Although he has come close, his number of exact identifications, or hits, among these three profiles would be zero, no better than it would be for some less astute rater who believed B, H, and J were the lowest three in the peer group on that characteristic. In short, the ''hits'' measure contains no provision for crediting near misses, but the correlation between the ranking of unidentified profiles and the ranking of his named peers on the success dimensions should provide a continuum which the raw ''hits'' metric does not possess. A rank-order correlation between these two ranks should provide a sensitive measure of recognition far more powerful than the simple count of matched profiles.

Data Analysis

In order to apply the metric described in the preceding paragraph, three rankings were collected. First, an official ranking (OR) of the students performed by the school was available. Second, a ranking of the anonymous profiles (UP) was collected. Finally, a ranking of seminar members by their peers (PR) was collected. This ranking was made using only a list of peer names, not profiles, and was made according to predictions of success in training.

The UP and PR rankings were group average ranks derived by summing all of the assigned ranks for each person in his seminar group, then converting that total sum of ranks back to a rank order ranging from 1 to 13 or 14 depending on the seminar's group size. These average ranks, UP and PR, represented a group concensus on the perception of each seminar member by the group. The Official Class Rank (OR) was determined by class standing on four exams (312 points), drill evaluation (25 points), student evaluation (25 points) and communication skills (38 points).

Rank-order correlations for each rater were computed for the following purposes:

(1) Correlation between unidentified profile ranking and named peer rankings (UP-PR)--one correlation coefficient was computed for each rater and was viewed as a more sensitive measure of hits than the number of exact identifications of unlabeled profiles. This produced a new variable, the logic of which was explained above.

(2) Correlations between unidentified profile rankings and official class rank (UP-OR)--One correlation coefficient for each rater. This variable indicates how well the rater can evaluate the operational criterion (OR) in terms of the statements available. Differences in effectiveness among the statement sets should be revealed in differences between the sizes of the average correlation coefficients. Average correlation coefficients across groups could have been computed by summing the numerators in the rho formula ($6\Sigma d^2$) and divided by the sum of the demoninators ($N(N^2 - 1)$). The squared deviations (d^2) were used in the analyses of variance since in this instance it provided a simpler and more accurate measurement variable in examining rank order effect than the correlation coefficients themselves.

(3) Correlation between named peer rankings and official class rank (PR-OR)--One for each rater. The average of this correlation coefficient would normally indicate the efficiency of peer ratings in predicting a criterion. In this case, however, there was considerable evidence that most of the subjects were well aware through intra-group discussion of how their peers had done on previous tests and were consequently aware of how they stood on the overall class evaluation. In short, they were ranking on direct information about their peers rather than judgment based on indirect knowledge.

The primary analysis included testing to see if significant differences existed in terms of hits and the other dependent variables among the three treatment conditions. Since each seminar group was randomly assigned to one of the three treatment conditions, the experimental design resulted in the nesting of three seminar groups under each treatment condition. The hierarchical design (Nested Factors) is usually used to test the effects among a number of treatments in certain types of experimental situations (Winer, 1962). Typical examples include investigating drug effects among a number of hospitals, studying teaching methods among a number of schools, or studying training methods among different individuals.

The hierarchical ANOVA is an efficient method of studying such experimental situations because it avoids multiple t-tests or non-orthogonal comparisons (Hays, 1963). The two-way hierarchical ANOVA in this experiment is also a more powerful statistical test than a one-way ANOVA that only tests for treatment effects ignoring any group effects. In this design, the nested factors are controlled by statistical procedures. In many experimental situations, it is dangerous to assume that certain nested factors have no significant influence on treatment effects.

Two sources of variation were observed in the experimental data. The treatment effect was of primary interest, whereas the seminar group affiliation was of secondary interest. The null hypothesis, no differences between treatment means, was tested for both investigated sources of variation. The analysis of both sources of variation was accomplished by performing a two-way hierarchical ANOVA for experiments with unequal cell sizes using the least-squares procedural method described by Tim and Carlson (1975).

''Hits'' and the sum of the squared differences between UP and PR rankings, UP and OR rankings, and PR and OR rankings were the dependent variables used in the ANOVA analysis to determine if significant differences existed among treatment conditions. The squared difference between rank orderings was used rather than the rank-order correlations since the squared difference provided a simpler and more accurate measurement variable in examining rank order similarity.

Results and Discussion

The hierarchical ANOVA summary for ''hits,'' or correct identification of profiles is shown in Table 1. As expected, the ''hit'' measurement variable showed no significant differences among treatments. In essence, the rating ''picture'' for each individual produced by the three different sets of rating statements were equal in their descriptive power. However, seminar group effects within a treatment were significant at the .01 level (Table 1). Table 2 shows the summary results of hits for seminar groups within treatments.

Table 1. Analysis of Variance by Number of ''Hits''
(Correct Profile Identifications) by
Treatment and Seminar Group

Source	Sum of Squares	D.F.	Mean Squares	F
Treatment	6.215	2	3.107	.349
Seminar Groups Within Treatments	53.421	6	3.903	3.247*
Error (Within Groups)	304.379	111	2.742	

*Significant at the .01 level.

Table 2. Number of Profile Identifications (Hits) by Treatment and by Seminar Group

	Treatment 1 (Seminar Group)			Treatment 2 (Seminar Group)			Treatment 3 (Seminar Group)		
	F	I	A	C	E	H	B	D	G
Group Results									
Total N	13	14	13	14	13	13	13	13	14
Total Hits	24	47	45	32	26	48	23	29	42
Mean Hits	1.86	3.36	3.46	2.29	2.00	3.69	1.77	2.23	3.00
SD Hits	1.63	1.82	2.37	1.90	1.68	1.55	1.30	1.30	.96
Treatment Results									
Total N		40			40			40	
Total Hits		116			106			94	
Mean Hits		2.90			2.65			2.35	
SD Hits		2.05			1.83			1.27	
T-Ratios									
Treatments 1 vs. 2 Comparison				$t = .574^{ns}$					
Treatments 1 vs. 3 Comparison				$t = 1.44^{ns}$					
Treatments 2 vs. 3 Comparison				$t = .85^{ns}$					

^{ns} = not significant.

The average rank-order correlations between the pairs of rankings appear in Table 3. Using Ferguson's (1966) table of significance for Spearman Rhos, 25 of the possible 27 rhos were significant at the .05 level. Furthermore, most of the nine correlations possible in each treatment group were significant at the .01 level (21 in all), and only one correlation in treatment II and III was not significant. All correlations demonstrated a similar pattern of significance in each of the three treatment conditions. The three rank order comparisons showed a high degree of agreement. This data analyses suggested that no one type of rating statement was superior for use in performance appraisal instruments. The purpose of these rank order comparisons was to see whether the pattern of significance under each treatment was generally similar or different. However, the most definitive test for determining differences between treatments was the hierarchical ANOVA analysis.

Tables 4 to 6 show the hierarchical ANOVA summary for comparison of the rating statement treatment conditions with respect to the squared difference between the following rank order comparisons: UP-PR, UP-OR, and PR-OR. The ANOVA results showed no significant difference between treatment conditions as reflected by the squared differences between the UP-PR rankings (viewed as a more sensitive measure of identification of unlabeled profiles), the UP-PR rankings (indicating how well the rater can evaluate the operational criterion in terms of given stimulus statements), and the PR-OR rankings (normally indicating the efficiency of peer ratings in predicting a criterion).

Table 3. Rank Order Correlations Among Unidentified Profile Rankings, Peer Rankings, and Official Rank by Treatment and by Seminar Group

	Treatments								
	I (Worker) Seminar Groups			II (Task) Seminar Groups			III (Trait) Seminar Groups		
Rank Order Comparisons	F	I	A	C	E	H	B	D	G
UP and PR	.58*	.86**	.87**	.85**	.86**	.90**	.79**	.90**	.71**
UP and OR	.52*	.71**	.82**	.43	.65*	.85**	.37	.72**	.70**
PR and OR	.87**	.93**	.97**	.57*	.79**	.94**	.74**	.79**	.97**
Total N	13	14	13	14	13	13	13	13	14

Critical values of p, the Spearman rank correlation, were obtained from Ferguson (1959), Table G, p. 414.
*Significant at .05 level.
**Significant at .01 level.

Table 4. Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Peer Rankings by Treatment and by Seminar Group

Source	Sum of Squares	D.F.	Mean Squares	F
Treatment	9396.114	2	4698.057	.117
Seminar Groups Within Treatments	241470.876	6	40245.146	4.722*
Error (Within Group)	945985.099	111	8522.388	

*Significant at .01 level.

Table 5. Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Official Rankings by Treatment and by Seminar Group

Source	Sum of Squares	D.F.	Mean Squares	F
Treatment	12127.327	2	6063.663	.0922
Seminar Groups Within Treatments	394330.700	6	65721.783	13.051*
Error (Within Groups)	558976.730	111	5035.826	

*Significant at .01 level.

Table 6. Analysis of Variance of Squared Deviations between Peer Rankings and Official Rankings by Treatment and by Seminar Group

Source	Sum of Squares	D.F.	Mean Squares	F
Treatments	119263.060	2	59631.530	.769
Seminar Groups Within Treatments	465196.015	6	77532.668	16.160*
Error (Within Groups)	53553.566	111	4797.780	

*Significant at .01 level.

The PR-OR rank order coefficient, however, cannot be considered an unbiased indicator since there was considerable evidence that most subjects were ranking on information based on knowledge of test performance acquired through intra-group association, rather than judgment based solely on observation of peer activities and traits.

Although no significant rank order differences were found between treatment conditions as reflected by the squared difference of the various pairs of rankings, the differences between seminar groups within treatments on all three ANOVA analyses were significant at the .01 level (Tables 4, 5, and 6). This was an unexpected finding because each seminar group was randomly assigned to one of the three treatment conditions. The results demonstrated that no one type of content rating statement was superior to any other in determining rank order differences.

The data analyses showed that the statements investigated here yielded no significant advantages for one set of statements over another. It makes no difference whether the rating statements are task-oriented, worker-oriented, or trait-oriented. This study provides additional evidence that the doubts of Bell, Hoff, and Hoyt (1963), Borman and Dunnette (1975), and Kavanagh, MacKinney, and Wollins (1971) about the superiority of job-oriented dimensions over trait-oriented dimensions were well founded. As Kavanagh (1971) concluded from his comprehensive literature review of performance appraisal studies, there is no reason to assume the superiority of job-oriented statements over trait-oriented statements. The selection of rating statements for inclusion in performance appraisal devices should primarily be determined by cost considerations. Cost considerations tend to favor trait-oriented statements in most situations, since job analysis, which is required to obtain task-oriented and worker-oriented statements, is costly and time consuming. Trait-oriented statements are also much more generalizable across different occupations than either task-oriented or worker-oriented statements.

Unlike many prior studies, this study does not conclude with a condemnation of judgmental rating statements. This study suggests that peer group person-oriented statements are as effective as job descriptive statements when the standard is an external criterion such as ability to recognize peers from unidentified profiles or ability to predict their official class rank.

An unexpected finding was the significant effect associated with seminar groups on all performed ANOVA analyses, particularly since all seminar groups were randomly assigned to each treatment condition. The importance of recognizing and controlling for group effects in such performance evaluation studies is evident. Investigated treatment variables might easily become contaminated by group effects leading to inaccurate results and conclusions. The reasons for these significant group effects are unknown, although such intra-group variables as morale, leadership, and attitude are possible causal influences.

It may be that performance appraisal research emphasis has not been placed on the most important variables. Perhaps there are environmental influences that affect performance ratings more than variables attributable to the appraisal device. Perhaps such issues as content, format, scale, etc., are relatively unimportant as compared to these other variables. A need exists to broaden the research focus in performance appraisal studies focusing on criteria independent and external to the performance appraisal device.

Summary and Conclusions

Three different kinds of rating stimulus statements differing along a dimension of trait-oriented to task-oriented descriptions, were compared in a context which permitted the comparisons to be made in terms of criteria external to the ratings. No evidence of superiority was found for any of the three sets although many significant correlations with various external criteria were obtained in all three experimental conditions.

Significant differences were also found among the three rating sub-groups comprising each of the three treatment groups although these rating sub-groups were assigned randomly to the three treatment groups. The importance of controlling for group effects in peer group studies was noted.

REFERENCES

Arvey, R. D., & Hoyle, J. C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. *Journal of Applied Psychology*, 1974, *59*, 61-68.

Barrett, R. S. *Performance rating* Chicago: Science Research Associates, 1966.

Bell, F. O., Hoff, A. L., & Hoyt, K. B. A comparison of three approaches to criterion measurement. *Journal of Applied Psychology*, 1963, *47*, 416-418.

Borman, W. C., & Dunnett, M. D. Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, 1975, *60*, 561-565.

Borman, W. C., & Vallon, W. R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology*, 1974, *59*, 197-206.

Burnaska, R. F., & Hollmann, T. D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology*, 1974, *59*, 307-312.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, *57*, 15-22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Curton, E. D., Ratliff, F. R., & Mullins, C. J. *Content analysis of rating criteria*. Proceedings of symposium on criterion development for job performance evaluation, June 23-24, 1977. In press.

Dickinson, T. L., & Tice, T. E. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, *9*, 421-438.

Dunnette, M. D. *Personnel selection and placement*. Belmont, CA: Wadsworth, 1966.

Hays, W. L. *Statistics for Psychologists*. New York: Holt, Rinehart, and Winston, 1963.

Kavanagh, M. J. The content issue in performance appraisal: A review. *Personnel Psychology*, 1971, *24*, 653-668.

Kavanagh, M. J., MacKinney, A. C., & Wollins, L. Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin*, 1971, *75*, 34-49.

Kelley, E. L., & Fiske, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: University of Michigan Press, 1951.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, 1976, *61*, 750-758.

McCormick, E. J., Finn, R. H., & Scheips, C. D. Patterns of job requirements. *Journal of Applied Psychology*, 1957, *41*, 358-364.

Ronan, W. E., & Prien, E. P. Perspectives on the measurement of human performance. New York: Appleton Century Crofts, 1971.

Schmidt, F. L., & Kaplan, L. B. Composite versus multiple criteria: A review and a resolution of the controversy. *Personnel Psychology*, 1971, *24*, 419-484.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: Approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, *47*, 149-155.

Timm, N. H., & Carlson, J. E. *Analysis of variance through full rank models*. Multivariate Behavioral Research Monograph No. 75-1. Published by the Society of Multivariate Experimental Psychology 1975.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 1972, *7*, 457-466.

Zedeck, S., & Blood, M. R. Foundations of behavioral science research in organizations. Monterey, CA: Brooks/Cole, 1974.

APPENDIX

WORKER-ORIENTED RATING DIMENSIONS

	Below Average	Average	Above Average	Well Above Average	Out-Average
1. Military appearance........	(A)	(B)	(C)	(D)	(E)
2. Participates in class activities.................	(A)	(B)	(C)	(D)	(E)
3. Communicates clearly by oral and written methods...	(A)	(B)	(C)	(D)	(E)
4. Amount of assistance to peers in work assignments..	(A)	(B)	(C)	(D)	(E)
5. Completes work in a timely manner.....................	(A)	(B)	(C)	(D)	(E)
6. Follows provided instructions...............	(A)	(B)	(C)	(D)	(E)
7. Takes accurate notes.......	(A)	(B)	(C)	(D)	(E)
8. Competence in analyzing work assignments...........	(A)	(B)	(C)	(D)	(E)
9. Awareness of safety precautions..............	(A)	(B)	(C)	(D)	(E)
10. Studies well on his own...	(A)	(B)	(C)	(D)	(E)

APPENDIX

TASK-ORIENTED RATING DIMENSIONS

		Below Average Effectiveness	Average Effectiveness	Above Average Effectiveness	Well Above Average Effectiveness	Outstanding Effectiveness
1.	Knows UCMJ programmed text.......	(A)	(B)	(C)	(D)	(E)
2.	Contributes examples in seminar on Discipline and Unity of Command............	(A)	(B)	(C)	(D)	(E)
3.	Promotes and organizes Community Project............	(A)	(B)	(C)	(D)	(E)
4.	Analyzes courts-martial case study.	(A)	(B)	(C)	(D)	(E)
5.	Participates in Foreign Policy role playing............	(A)	(B)	(C)	(D)	(E)
6.	Understands reasons for nonalignment of uncommitted nations	(A)	(B)	(C)	(D)	(E)
7.	Knows history of AF uniform.........	(A)	(B)	(C)	(D)	(E)
8.	Applies the six-step approach to problem solving............	(A)	(B)	(C)	(D)	(E)
9.	Knows how to plan a conference.........	(A)	(B)	(C)	(D)	(E)
10.	Researches topic for Persuasive Speech..	(A)	(B)	(C)	(D)	(E)

APPENDIX

TRAIT-ORIENTED RATING DIMENSIONS

	Below Average	Average	Above Average	Well Above Average	Out-standing
1. Honesty - straightforward and truthful in dealing with others	(A)	(B)	(C)	(D)	(E)
2. Ambition - works hard, accepts challenges	(A)	(B)	(C)	(D)	(E)
3. Dependability - does assigned tasks conscientiously without close supervision	(A)	(B)	(C)	(D)	(E)
4. Punctuality - prompt in keeping engagements	(A)	(B)	(C)	(D)	(E)
5. Quality of work - performs work accurately and effectively	(A)	(B)	(C)	(D)	(E)
6. Quantity of work - produces a large amount of work that meets requirement standards	(A)	(B)	(C)	(D)	(E)
7. Initiative - originates and achieves goals on his own	(A)	(B)	(C)	(D)	(E)
8. Adaptability - changes attitude and behavior to meet the demands of the situation	(A)	(B)	(C)	(D)	(E)
9. Originality - creative, thinks of new solutions to old problems	(A)	(B)	(C)	(D)	(E)
10. Agreeableness - gets along well with fellow workers, well liked	(A)	(B)	(C)	(D)	(E)

DIFFERENTIAL RESPONSES ON ALTERNATELY ANCHORED JOB RATING SCALES[1]

Jimmy L. Mitchell, Lt Col, USAF

USAF OCCUPATIONAL MEASUREMENT CENTER
OCCUPATIONAL SURVEY BRANCH
LACKLAND AFB, TEXAS 78236

A paper presented at the Military Testing Association Convention

30 October - 3 November 1978

---

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

DIFFERENTIAL RESPONSES ON ALTERNATELY ANCHORED JOB RATING SCALES

Jimmy L. Mitchell, Lt Col, USAF

USAF Occupational Measurement Center
Occupational Survey Branch
Lackland AFB, TX 78236

A variety of rating scales have been used with job and occupational data through the years but very seldom is a rationale given for the use of a particular scale. Likewise, there have been a number of ways in which scales have been anchored but the reasons behind the choice of a 5point scale over a 7- or 9-point scale have not typically been reported.

Viteles' job psychograph was developed in 1934; it consisted of a standard set of psychological traits, each of which was to be rated by a job analyst as to its "importance" for the job being studied (Viteles; as cited in Blum & Naylor 1968; 506). The considerable influence of this pioneering work survives today in the form of trait ratings, such as are used in the Department of Labor job analysis system (Department of Labor 1972) and in the wide-spread use of 5-point importance scales (cf. Baehr 1967; McCormick, Jeanneret, & Mecham 1972).

In some of the more recently developed job analysis systems, longer scales have been used. Hemphill (1959) in his study of executive positions, used a 7-point Part-of-the-Position scale with three verbal anchors. The Air Force occupational analysis program used first a 7-point scale and later a 9-point scale measuring relative time spent, with verbal anchors for each scale point (Morsh 1964, Driskill 1975). Other job analysis systems have used scales which vary in length from item to item (Scott 1963; Fine and Wiley 1971).

The literature on scaling provides few clues as to the optimum number of levels for job rating scales. However, Matell and Jacoby (1972) determined experimentally that if the number of scale levels exceeds 5, only about 60 percent of the scale will be used. They concluded that scales of no more than five to seven levels should be adequate for most measurement purposes.

Christal and Madden (1961) have raised the issue of being able to detect those jobs which would be "off scale" when compared to other jobs. This is an issue of particular interest when a large number of jobs are to be considered and one objective of measurement is to be able to distinguish between jobs which are substantially different. In such cases, a larger number of scale levels are needed to insure that the extreme jobs can be appropriately rated. Thus, in the Air Force occupational analysis program, a 9-level scale is typically used. This gives the maximum possible discrimination in a single digit scale and provides the opportunity to detect extreme jobs in most Air Force occupational areas.

A potentially more serious problem lies in the selection of verbal anchors for the scale points of job rating scales. Christal and Madden (1961) noted that it has never been determined whether every scale point should have a verbal anchor. While most job rating scales which have been used through the years have provided such anchors, Hemphill (1959) used a 7-point scale with only three verbal anchors. Cragun and McCormick (1967) used this scale with Air Force officers in a study of the reliability of job ratings and their results suggest that it had considerable reliability and was to some degree preferred by incumbents in managerial positions to characterize their jobs. Tornow and Pinto (1976) used the same scale but they compressed it to a five point scale; they provided no rationale for their modification of the Hemphill scale nor any estimate of the effect of this modification on their final data.

I have not been able to find any definitive answer to the question of the anchoring of scale points in the job analysis literature. However, in the course of gathering and analyzing data for the development of a structured job analysis instrument, I chanced on some interesting results which bear on this issue.

The instrument being developed was the Professional and Managerial Position Questionnaire (PMPQ), an experimental structured job analysis questionnaire for the study of higher level jobs (Mitchell and McCormick 1976). This 93-item questionnaire was developed in the tradition of McCormick's Position Analysis Questionnaire (PAQ) but was aimed specifically at executive and management types of positions since earlier research with the PAQ had indicated that a separate instrument for higher-level positions might be appropriate (Harris & McCormick 1973).

In this new instrument, 9-point Part-of-the-Job and Complexity scales were used with verbal anchors for every other scale point (1, 3, 5, 7, and 9). Additionally, the Complexity ratings were further anchored with behavioral examples; these behavioral examples were scaled by obtaining independent ratings of a set of examples from panels of professional and academic industrial psychologists (Mitchell 1978). Also included in the instrument were items dealing with the personal requirements for the positions, to determine such things as educational levels required, prior experience, training, etc., and a section for other information, such as the number of people supervised, etc. For these items, there were numbers, categories, or constructs which were used to anchor every point of the scale, such as years of education, numbers of employees, etc. Thus, in the same instrument, there were both alternately anchored items (Part-of-the-Job, Complexity) and items with verbal anchors for every scale point (Number supervised, etc.).

The PMPQ was used to gather data on 300 positions in 45 companies, schools, and government agencies throughout the country. The sample of jobs was quite diverse and salary levels ranted from about $690 per month for an administrative assistant to over $6800 per month for an executive vice president of a major company. About 250 cases had complete data and were useable in the various types of analysis planned for the study. An analysis of the distribution of responses by item was not included in the research plan but in the course of displaying some of the data for another purpose, it was noted that some items appeared to have non-normal distributions. This led to displaying the data in such a way that the distribution of responses by scale point was visible. Table 1 gives a partial picture of this data.

The items at the top of this table are those with alternately anchored response categories. Items at the bottom have a verbal anchor for each scale point.

TABLE 1

RESPONSE DISTRIBUTIONS FOR A SAMPLE OF ITEMS FROM THE PMPQ

ITEMS		RESPONSES									
		0	1	2	3	4	5	6	7	8	9
1.	Work Scheduling (P)	5	7	2	27	12	74	15	65	18	29
2.	Complexity of Work Scheduling (C)	4	10	4	52	26	85	29	26	8	7
43.	Planning/Scheduling (Summary P)	2	6	3	32	19	85	21	55	10	26
	* * *										
67.	Formal Education Required	0	4	15	15	27	124	6	36	12	36
87.	No. of Nonsupervisory Personnel	55	45	53	28	17	11	15	16	10	10
89.	Total No. of Personnel	26	37	56	36	28	16	31	13	6	1

You will note that for the alternately anchored items, the 2, 4, 6, an 8 response categories are consistently lower than are the 1, 3, 5, 7 an 9 categories. This pattern is perhaps even more visible if the data are plotted as histograms.

Figure 1 gives the distribution of responses for Item 1, which asks the degree to which an incumbent schedules his or her own work or the work of others. The verbally anchored scale points are indicated in this figure by cross hatching while the unanchored scale points are shown blank. You can see that all response categories were used but that there is a marked differential in response between the anchored and the unanchored scale points.

Figure 2 displays the distribution of responses for the second item in the PMPQ; how complex are the work scheduling activities of the position? Here, the anchored scale points have not only a verbal anchor but also have one or two behavioral examples to concretely reference the level of complexity. Again, there is a marked differential in response frequency between anchored and unanchored response categories.

Figure 3 represents data from Item 89, which asks the total number of personnel in units under the supervision or management control of the incumbent. Here all response categories are concretely anchored with an interval; for this item, 3 = 10 to 25 people. As you can see from the distribution of responses displayed in this figure, this is quite a different kind of distribution. There is no marked difference across adjacent items in the systematic way seen in Figures 1 and 2. Thus, there appear to be very major differences in the way individuals respond to anchored and unanchored rating scales.

We have not yet tested to see if these are significant differences. Hopefully, this work can be done in the next few months and we can come to a more concrete conclusion. When this is done, I expect that we will seek to publish the result as a short note in one of the journals.

For the present, this unexpected result has led me to question the results of some of the earlier research. Would the results of Hemphill's landmark study of executive positions have been the same had he used a verbal anchor for all scale points rather than just three anchors across seven response categories? Would Cragun and McCormick have come to the same conclusions if they had used a Part-of-the-Postion scale which was completely anchored? Of course, there are no ready answers to these questions. We have not yet done the research needed to clarify just what is going on in these cases nor do we yet have any idea of the impact of this differential response phenonomenon on the major findings of earlier research.

What is clear is that this is a phenomemon which must be looked into; we need to learn how this type of differential response tendency impacts on occupational data and ultimately on management decisions made with these data.

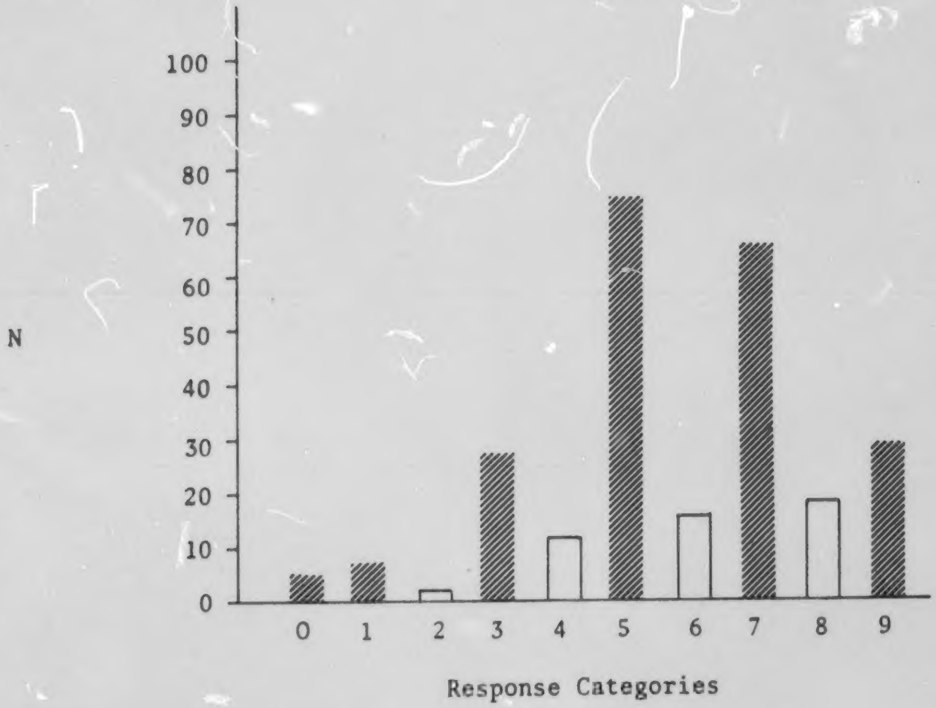


Figure 1. Distribution of responses from PMPQ Item 1. - Work Scheduling (P)

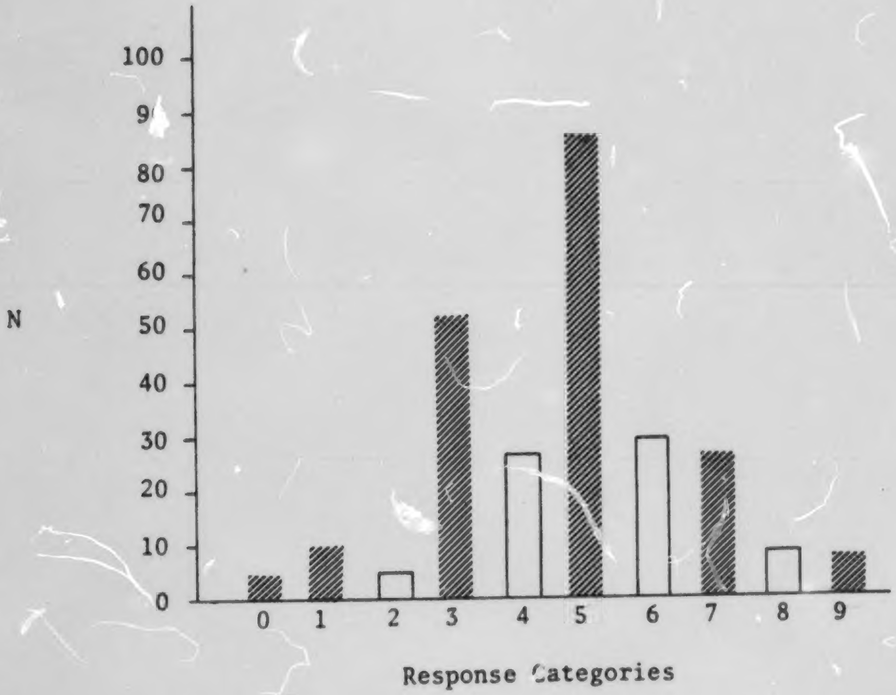


Figure 2. Distribution of responses for Item 2. - Complexity of Scheduling

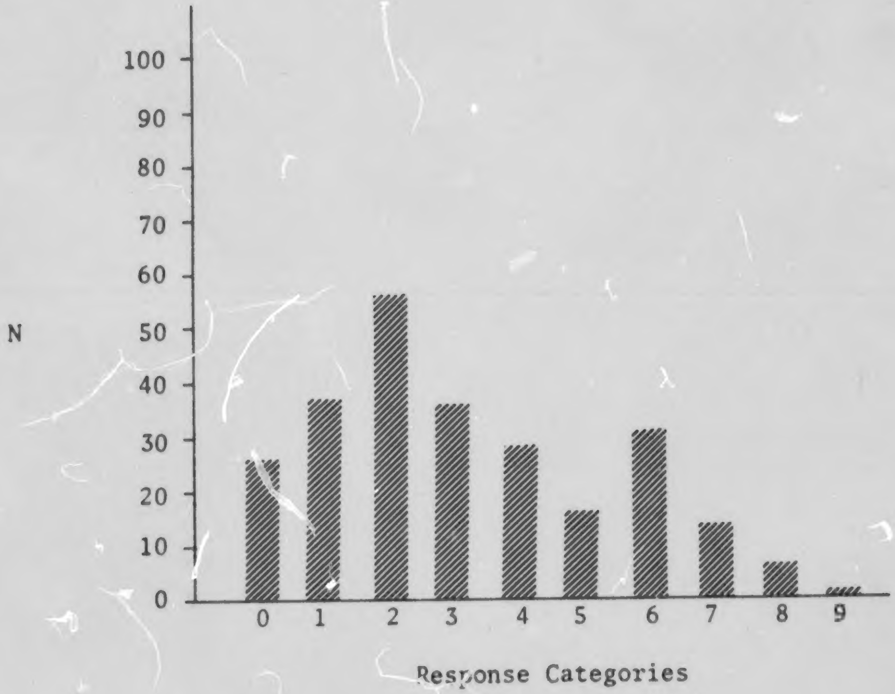


Figure 3. Distribution of responses for Item 89. - Total Number of Personnel in Units Supervised

For the present, we must assume that this type of differential response is not a desirable outcome and thus, that alternately scaled items should be avoided. Until more is known about the impact of variance in verbal anchoring such scales, scales with verbal anchors for each response category should be used. If verbal anchors cannot be developed for each scale point, then we perhaps should use a semantic differential with anchors only at the end points. It would be interesting indeed to see how our results would vary with these different anchoring systems this is an area which really could benefit from some empirical research.

BIBLIOGRAPHY

Baehr, M. E. A factorial framework for job descriptions for higher-level personnel. Industrial Relations Center, The University of Chicago, 1967.

Christal, R. E., and J. M. Madden. Air Force research on job evaluation procedures. ASD-TN-61-46. Lackland AFB, TX: Personnel Laboratory, June 1961.

Cragun, J. R., and E. J. McCormick, Job Inventory information; task and scale reliabilities and scale interrelationships. PRL-TR-67-15, Lackland AFB, TX: Personnel Research Laboratory, November 1967.

Depart of Labor. Handbook for Analyzing Jobs. Washington, D. C.: Government Printing Office, 1972.

Driskill, Walter E. Occupational Analysis in the United States Air Force. Paper presented at the Task Inventory Exchange National Symposium on Task Analysis/Inventories, The Ohio State University, Columbus, Ohio, November 18, 1975.

Fine, S. A., and W. W. Wiley. An Introduction to Functional Job Analysis. Kalamazoo, MI: The W. E. Upjohn Institute for Employment Research, 1971.

Harris, A. F. and E. J. McCormick. The analysis of rates of Naval compensation by use of a structured job analysis procedure. Report No. 3. Purdue University, Department of Psychological Sciences, September 1973.

Hemphill, J. K. Job descriptions for executives. Harvard Business Review, 1959, 37, 55-67.

Matell, M. S. and J. Jacoby. It there an optimal number of alternatives for Likert-scale items? Journal of Applied Psychology, 1972, 56 506-509.

McCormick, E. J., P. R. Jeanneret, and R. C. Mecham. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ), Journal of Applied Psychology Monograph, 1972, 56 347-368.

Mitchell, J. L. Structured job analysis of professional and managerial positions. Unpublished doctoral dissertation, Department of Psychological Sciences, Purdue University, West Lafayette, IN, 1978.

Mitchell, J. L. and E. J. McCormick. Professional and Managerial Position Questionnaire. Department of Psychological Sciences, Purdue University, West Lafayette, IN, 1976.

Morsh, J. E., Job analysis in the United States Air Force. Personnel Psychology 1964, 17, 7-17.

Scott, W. E., Jr. The reliability and validity of a six-factor job evaluation system. Unpublished doctoral thesis, Department of Psychological Sciences, Purdue University, West Lafayette, IN, 1963.

Tornow, W. W., and P. R. Pinto. The development of a managerial job taxonomy a system for describing, classifying, and evaluating executive positions. Journal of Applied Psychology, 1976, 61, 410-418.

Viteltes, M. S. A psychologist looks at job evaluation. Personnel, 1941, 16, 165-176, as cited in Blum and Naylor, Industrial Psychology, New York: Harper & Row 1968.

SAMPLE SIZE AND STABILITY OF TASK ANALYSIS INVENTORY RESPONSE SCALES

John J. Pass and David W. Robertson
Navy Personnel Research and Development Center
San Diego, California 92152

Problem

Occupational task analysis inventories are administered on a recurring basis to hundreds of thousands of personnel in the military services. While the collected data are used by management for the specification of occupational standards, the design of training curricula, and the structuring of occupational specialities, the data acquisition procedures place heavy time demands on job incumbents. Typical inventories contain between 800 to 1000 items and can take over four hours to administer. Thus, the problem is, how to minimize the time demands on the Fleet while selecting sample sizes and inventory response scales adequate to obtain stable (that is, reliable) data.

Objective

The objective of the study was to determine empirically the stability and independence of responses on two task analysis response scales--the Time-Spent scale and the Task-Performed scale (these scales are currently in use by the military services--they will be defined subsequently). A primary concern was the degree of change in stability as sample size varied.

METHOD

Data

Task inventory response data (Display 1) were provided by the Navy Occupational Development and Analysis Center (NODAC). Four Navy occupational specialties (termed Ratings in the Navy) were selected for analysis; that is, the Aviation Machinist's Mate, the Electronics

---

Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, 30 October to 3 November, 1978.

The opinions and assertions contained herein are those of the writers and are not to be construed as official or reflecting the views of the Navy Department.

Technician, the Torpedoman's Mate, and the Yeoman. These four occupations were deemed to be representative of a broad range of occupational requirements. The data were collected from job incumbents in a wide variety of both Fleet and Shore activities.

Each of the four data sets provided by NODAC was randomly split by paygrade to obtain eight pairs of independent paygrade samples (for paygrades E2 to E9), each comprising 50 percent of the paygrade personnel in the respective total sample.

Inventory Scale Response Data

The fundamental task analysis data collected by the military occupational analysis programs are responses to the Time-Spent scale (a scale developed within the Personnel Research Laboratory of the U.S. Air Force). This scale is a Likert-type scale of time spent performing a task, with scale points ranging from "very much" through "average" to "very little." The Navy program uses a five point Time-Spent scale. Other military services use a seven or nine point Time-Spent scale.

The Comprehensive Occupational Data Analysis Programs (CODAP), a programming package developed and upgraded by personnel of the Human Resources Laboratory of the Air Force, operates on the Time-Spent responses and converts these data, as shown in Display 2, to responses on a Relative Time-Spent scale and to responses on a binary Task-Performed scale (where a score of 1 indicates the task is performed and 0 indicates the task is not performed). From these converted response scores, average scores for a given sample are derived by CODAP for the tasks in an inventory. These average score vectors, called job descriptions or job description profiles (Display 3), contain the most widely used task analysis information. As shown in Display 3, the first vector or profile is the percent of personnel performing each task, calculated by taking the average of responses on the Task-Performed scale. The other two profiles are averages of responses on the Relative Time-Spent scale. The profile in the middle of the Display is calculated on scores for only those personnel who perform the task; that is, personnel with zero or blank Time-Spent scores are not included in the calculation of these averages. All the personnel are included in the calculation of average percentages of Time-Spent for the third profile shown. The data in this display are actual data derived from scale responses by paygrade 6 personnel from the Torpedoman's Mate Rating.

The present study derived these three profiles for the randomly drawn independent paygrade samples, and calculated the similarity between each profile, based on several indices, across samples (Display 4). Of primary interest was the degree of similarity of the profile data for corresponding paygrade samples in each rating as indicated by the X's in the diagonal of the display matrix. Since the job description profiles are averages of responses on either the Relative Time-Spent or the Task-Performed scales, the degree of obtained similarity between corresponding paygrades indicates the degree of stability of the responses on these scales.

Stability Indices

Three stability indices were calculated on the profile data (Display 5). All three reflect the stability over all tasks in the profile or inventory. For the Product Moment (PM) coefficient calculation, profile tasks were treated as cases, and percentages as scores.[1] Essentially this coefficient measured the stability of the relative values or rank order of inventory tasks in terms of Relative Time-Spent or Task-Performed percentages.

The other two indices measured the stability of the absolute or actual percentage values for the percent performing profile only. These indices evaluated the difference in percentages of personnel performing the same tasks across independent paygrade samples. The percentage of inventory tasks that met the criteria listed on Display 5, that is, not exceeding 5 or 10 percentage points difference or not reaching significance, was the value for the particular index. Pairs of zero scores on corresponding tasks across samples were not included in the calculation of any index. The obtained values for certain of these indices were then plotted against sample size, and eta coefficients were calculated to measure the relationships. A computerized curve smoothing procedure (ISSC, 1970, pp. 11-7 to 11-9) was applied to the plots.

Independence of Responses to the Task-Performed and Time-Spent Scales

Using the same correlational model previously described, the Product Moment coefficient was also calculated between the Percent Performing profile and the Average Time-Spent by All personnel profile. This analysis was performed between these two profiles since preliminary results showed marked similarity; that is, a lack of independence between these profile data.

---

[1] With this correlational model, complete independence of scores did not exist. That is, the same individuals provided responses for calculation of a percentage (i.e., score) for more than one task. However, Cragun and McCormick (1967) report only minor inflation for coefficients derived with this same model for the study of U.S. Air Force task analysis inventory reliability.

583

RESULTS AND DISCUSSION

For this presentation, only some of the results will be presented. A technical report which includes all results related to this presentation and which also includes findings on the relationship between sample size and cluster solution stability is in preparation.

Comparative Stability and Independence of Responses on the Task-Performed and Time-Spent Scales

The stability results based on the PM coefficient are presented in Display 6. As shown, two of the profiles obtained very high median coefficients, but the profile, calculated on Relative Time-Spent values for only those personnel who perform the tasks, obtained relatively low stability values. The coefficient values for the other two profiles (i.e., Percent Performing and Average Time-Spent by All) were not only very high but also appeared to be positively related.

This apparent relationship was investigated to determine the degree of independence between profile data (Display 7). The very high correlation coefficients obtained between the Percent Performing and Average Time-Spent by All profiles within each of eight paygrades for AD and TM are shown in Display 7. Similar findings on the lack of independence between these two profiles are reported in a report published by the Human Resources Laboratory of the U.S. Air Force (see Carpenter, 1974). Thus, there is little difference between these profiles in terms of distributional shape or rank order of task scores (see Cronbach and Gleser, 1953). The use of either profile in determining rank order of tasks will yield very similar results.

Next the magnitude of these two similar profiles was examined. The percentages for the average Time-Spent by All profile are extremely small in value. The data in Display 3 are sorted from high to low on the basis of this profile's values. As shown, 1.98 percent is the largest score for the 337 TM inventory tasks for this sample of paygrade 6 personnel. Typically, values on this profile for all paygrade samples analyzed were below 1 percent, that is, an average of less than 1 percent of time was spent performing any task. The magnitude of these values make interpretation difficult. Parenthetically, small values were also typical for the other Time-Spent profile. Furthermore, Navy users surveyed reported little or no use of the Time-Spent data. On the other hand, the percentages of personnel performing tasks appear meaningful as well as being highly stable.

Other studies indicate additional problems with Time-Spent data; specifically, a less favorable reaction by job incumbents to using the Time-Spent scale as compared to other task analysis scales (see Cragun and McCormick, 1967), a substantial amount of time needed to mark tasks on the Time-Spent scale (estimated to be about 2.5 hours for 450 items out of the 800 to 1000 items in a typical inventory [Cragun and McCormick, 1967]), as well as inconsistent conclusions drawn in regard to the scale's validity (see Hartley, Brecht, Pagery, Weeks, Chapinis, and Hoecker, 1977 versus Carpenter, Giorgia, and McFarland, 1975; also see McCormick, in Dunnette, 1976, p. 670). Hartley et al. (1977) do report substantially valid rank ordering of tasks by job incumbents in terms of time spent. In comparison to the Time-Spent data, the Percent Performing profile based on Task-Performed data is highly stable and is used regularly by consumers of task analysis information. Thus, these data were selected to examine in relation to sample size.

Sample Size and Stability of Responses to the Task-Performed Scale

Display 8 plots the relationship between the correlational stability index calculated on the Percent Performing data against sample size. For comparability with other plots, correlation values were multiplied by 100 before plotting. As stated before, this index reflects the stability of the rank order of tasks for these Task-Performed data. The clearly asymptotic curve indicates high stability of data for sample size exceeding about 30 and extremely high stability when the sample exceeded about 100. This curve shows minimal improvement in stability for increases in sample size above about 40.

Display 9 shows two curves which plot the percentage of inventory tasks that did not exceed a difference across samples of either 10 or 5 percentage points. Curve 1 is clearly asymptotic and indicates high stability for sample size exceeding about 30 and extremely high stability when the sample exceeded about 100. Curve 2, reflecting the more rigorous criterion level, indicates very high stability at $\underline{N}$ above 240, and moderate stability at sample size above 100. The eta coefficients were .76 and .88 ($P \leq .01$, df=5, 26, see Hays, 1963, formula 16.6.4) for Curve 1 and 2, respectively, which indicate a substantial, highly significant relationship between sample size and stability.

Examination of the curves in relation to each other reveals additional information. First of all, the curve based on the correlational index is highly similar to the curve based on the 10 percent level. Thus, for interpretations of the data for sample size above about 40, Curve 1 in Display 9 can be considered to also represent the curve in Display 8 based on the correlational index.

Curve 2 in Display 9 intersects a stability value of about 75 (that is, 75 percent of inventory tasks across samples differed by less than 5 percentage points) at sample size of about 100. The question as to the stability (or amount of difference obtained) for the remaining 25 percent of inventory tasks is answered by examining the value at which Curve 1 intersects the stability dimension for the same sample size of 100. The value shown is about 97 percent and indicates that of the remaining 25 percent of inventory tasks, all but 3 percent differed by 10 or less percentage points. For another example of information gained by comparing curves, at sample size of 80, Curve 2 intersects the stability dimension at a moderate score of about 70, but Curve 1, when considered the same as the curve based on the correlation index (Display 8), intersects at about 95 (that is, a correlation coefficient equal to about .95). Thus, for this sample size of 80 the relative values or the rank order of all the tasks in the inventory are (is) highly stable in terms of percentages of personnel performing those tasks.

Display 10 shows data from all three curves. As shown, sampling beyond an N of 240 would produce very little gain even in terms of the most rigorous stability criterion. And if only the rank order of tasks in terms of numbers of people performing them is required, a sample size of 100 or even 40 would be acceptable. Consideration of available personnel and the information displayed resulted in a recommended total sample size of about 1400, or 45 percent less personnel than in the collected data for the AD Rating. A similar sample size was indicated for the ET Rating; that is, a sample containing about 1000 less personnel than in the existing total sample. Examination of the total sample sizes for about 36 Ratings reveals some oversampling for about one-fourth of the Ratings. On the other hand, an additional 115 personnel to add to the total sample of TM personnel was indicated by the findings. The application of these guidelines will enable more cost-effective sampling (especially realized for the larger Rating populations) and assure overall stability of results.

It should be noted that the utility of these obtained relationships relies on the degree of representativeness of the samples analyzed and those to be inventoried. Assuring a representative sample could require increasing sample size above that indicated by the study's guidelines. Other factors such as availability of personnel, and subgroups of special interest, must also be considered in determining sample size. One other possible limitation concerns the generality of these findings to other Ratings and to other types of occupational specialties. It is reasonable to expect the findings to apply to occupational specialties judged to be as homogeneous as (or more homogeneous than) paygrades within a Rating.

CONCLUSIONS

Based on the study's findings and current task analysis procedures, it is concluded that (Display 11):

1. To substantially reduce administration time, the Time-Spent scale can be deleted from future task analysis inventories without loss of practical information. Alternate methods of estimating time spent, including incumbent ranking of the most time consuming tasks, could be administered on a trial basis.

2. Responses to currently administered inventory scales could be used to calculate the percentage of incumbents performing tasks--CODAP modification is not essential.

3. The study's empirically developed guidelines on sample size required for stable data can be used an an aid to determine cost-effective sample sizes that optimize stability.

597

REFERENCES

Carpenter, J. B. *Sensitivity of Group Job Descriptions to Possible Inaccuracies in Individual Job Descriptions*. San Antonio: Air Force Human Resources Laboratory, March 1974. Technical Report AFHRL-TR-74-6.

Carpenter, J. B., Giorgia, M. J., & McFarland, B. P. *Comparative Analysis of the Relative Validity for Subjective Time Rating Scales*. San Antonio: Air Force Human Resources Laboratory, December 1975. Technical Report AFHRL-TR-75-63.

Cragun, J. R., & McCormick, E. J. *Job Inventory Information: Task and Scale Reliabilities and Scale Interrelationships*. San Antonio: Personnel Research Laboratory, November 1967. Technical Report PRL-TR-67-15.

Cronbach, L. J., & Gleser, G. C. Assessing Similarity between Profiles. *The Psychological Bulletin*, 1957, *50*, No. 6, 456-473.

Hartley, C., Brecht, M., Pagery, P., Weeks, G., Chapinis, A., & Hoecker, D. Subjective Time Estimates of Work Tasks by Office Workers. *Journal of Occupational Psychology*, 1977, *50*, 23-36.

Hays, W. L. *Statistics*. New York: Holt, Rinehart and Winston, Inc., 1963.

Integrated Software Systems Corporation (ISSC). *DISSPLA Plotting System, Intermediate Manual* (3rd Ed). San Diego, 1973.

McCormick, E. J. Job and Task Analysis. In Dunnette , M.D. (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally, 1976.

DISPLAY 1
TASK ANALYSIS SURVEYS FOR FOUR NAVY RATINGS ANALYZED

RATING		INVENTORY SIZE		
TITLE	ABBREVIATION	TOTAL ITEMS	TASKS	TOTAL SAMPLE SIZE
AVIATION MACHINIST'S MATE	AD	1163	404	2538
ELECTRONICS TECHNICIAN	ET	1080	597	2548
TORPEDOMAN'S MATE	TM	782	337	735
YEOMAN	YN	810	529	2771

DISPLAY 2

FUNDAMENTAL TASK ANALYSIS DATA: RESPONSES TO THE TIME-SPENT SCALE

	Job Incumbent		
Task	Time-Spent Response	Relative Time-Spent Response (%)	Task-Performed Response
1	0	0	0
2	2	20	1
3	5	50	1
4	0	0	0
5	3	30	1
	10	100%	

DISPLAY 3

DATA ANALYZED: CODAP JOB DESCRIPTION PROFILES DERIVED FROM RESPONSES ON THE RELATIVE TIME-SPENT AND TASK-PERFORMED SCALES

	JOB DESCRIPTION PROFILE		
TASK	PERCENT PERFORMING (%)	AVERAGE TIME-SPENT (%)	AVERAGE TIME-SPENT BY ALL (%)
1	82.42	2.41	1.98
2	90.11	1.95	1.76
3	74.72	1.96	1.46
4	72.52	1.87	1.35
5	67.03	1.90	1.28
6	63.74	1.77	1.13
7	61.54	1.59	.98
8	64.83	1.50	.97
9	63.74	1.51	.96
10	59.34	1.60	.94

DISPLAY 4
DETERMINATION OF STABILITY BASED ON COMPARISONS OF JOB DESCRIPTION PROFILES DERIVED FOR INDEPENDENT PAYGRADE SAMPLES

Sample B	Sample A Paygrade							
Paygrade	E2	E3	E4	E5	E6	E7	E8	E9
E2	X							
E3		X						
E4			X					
E5				X				
E6					X			
E7						X		
E8							X	
E9								X

DISPLAY 5
STABILITY INDICES CALCULATED ON JOB DESCRIPTION PROFILES ACROSS INDEPENDENT PAYGRADE SAMPLES

RELATIVE VALUE (RANK ORDER) STABILITY

1. PM CORRELATION COEFFICIENT

ABSOLUTE (ACTUAL) VALUE STABILITY

1. PERCENTAGE OF CORRESPONDING PROFILE TASKS THAT DO NOT EXCEED:
   A. 5 PERCENT DIFFERENCE
   B. 10 PERCENT DIFFERENCE
2. PERCENTAGE OF CORRESPONDING PROFILE TASKS THAT ARE NOT SIGNIFICANTLY DIFFERENT (Z TEST)

DISPLAY 6
COMPARATIVE STABILITY OF JOB DESCRIPTION PROFILES BASED ON PM CORRELATION COEFFICIENT

	Median Correlation Coefficient Across Corresponding Paygrades (E2-E9)		
Rating	Percent Performing	Average Time-Spent	Average Time-Spent By All
AD	.98	.33	.96
ET	.98	.50	.96
TM	.90	.32	.92
YN	.97	.31	.96

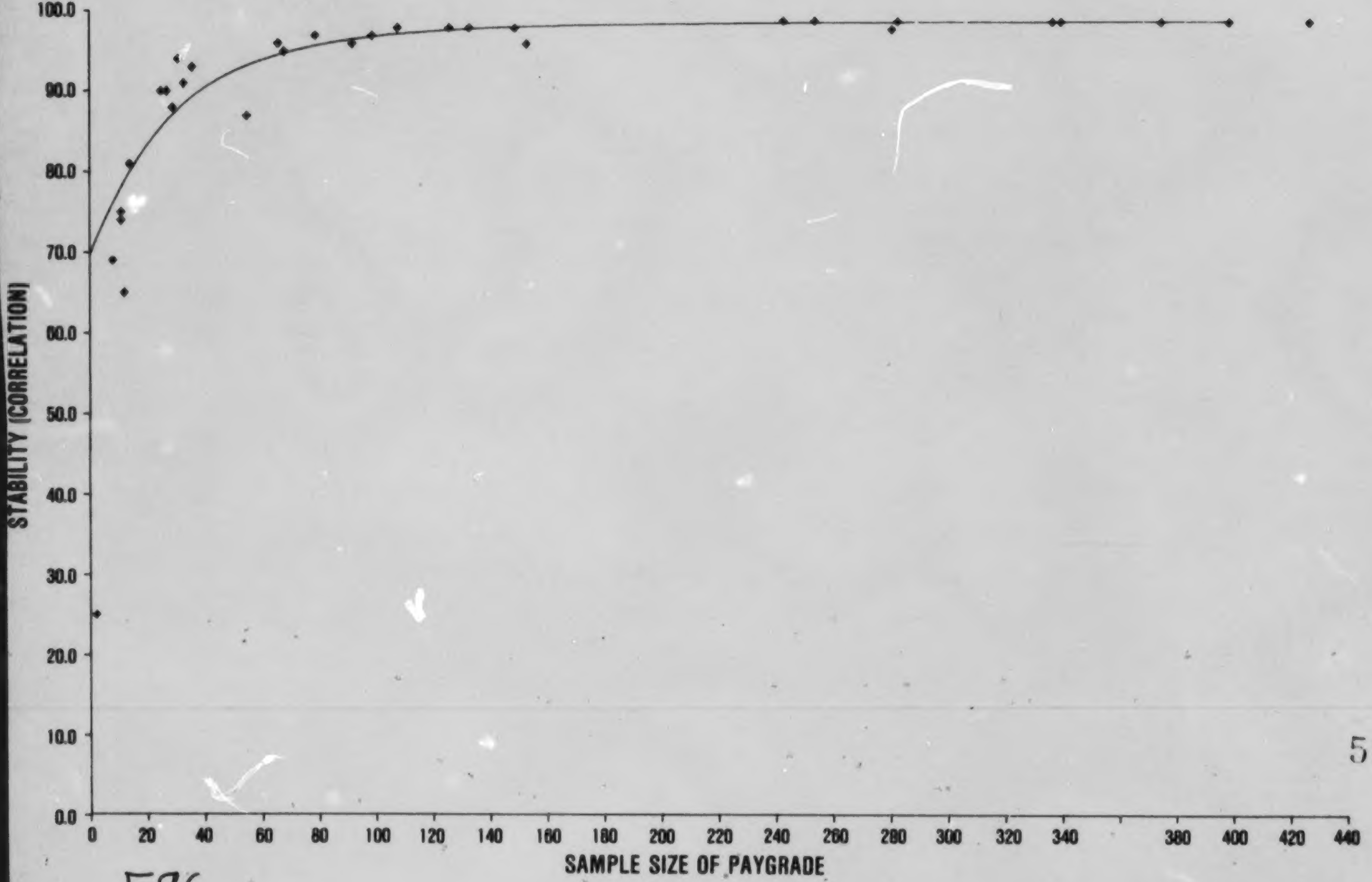
DISPLAY 7

PM CORRELATION BY PAYGRADE BETWEEN PERCENT PERFORMING AND AVERAGE TIME-SPENT BY ALL PERSONNEL PROFILES

	PAYGRADE							
RATING	E2	E3	E4	E5	E6	E7	E8	E9
AD	94 (67)	96 (149)	97 (282)	97 (337)	93 (281)	96 (108)	96 (31)	72 (14)
TM	92 (08)	94 (29)	96 (66)	96 (125)	94 (92)	90 (36)	92 (10)	83 (02)

NOTE: NUMBER IN PARENTHESIS IS NUMBER OF PERSONNEL IN PAYGRADE SAMPLE.

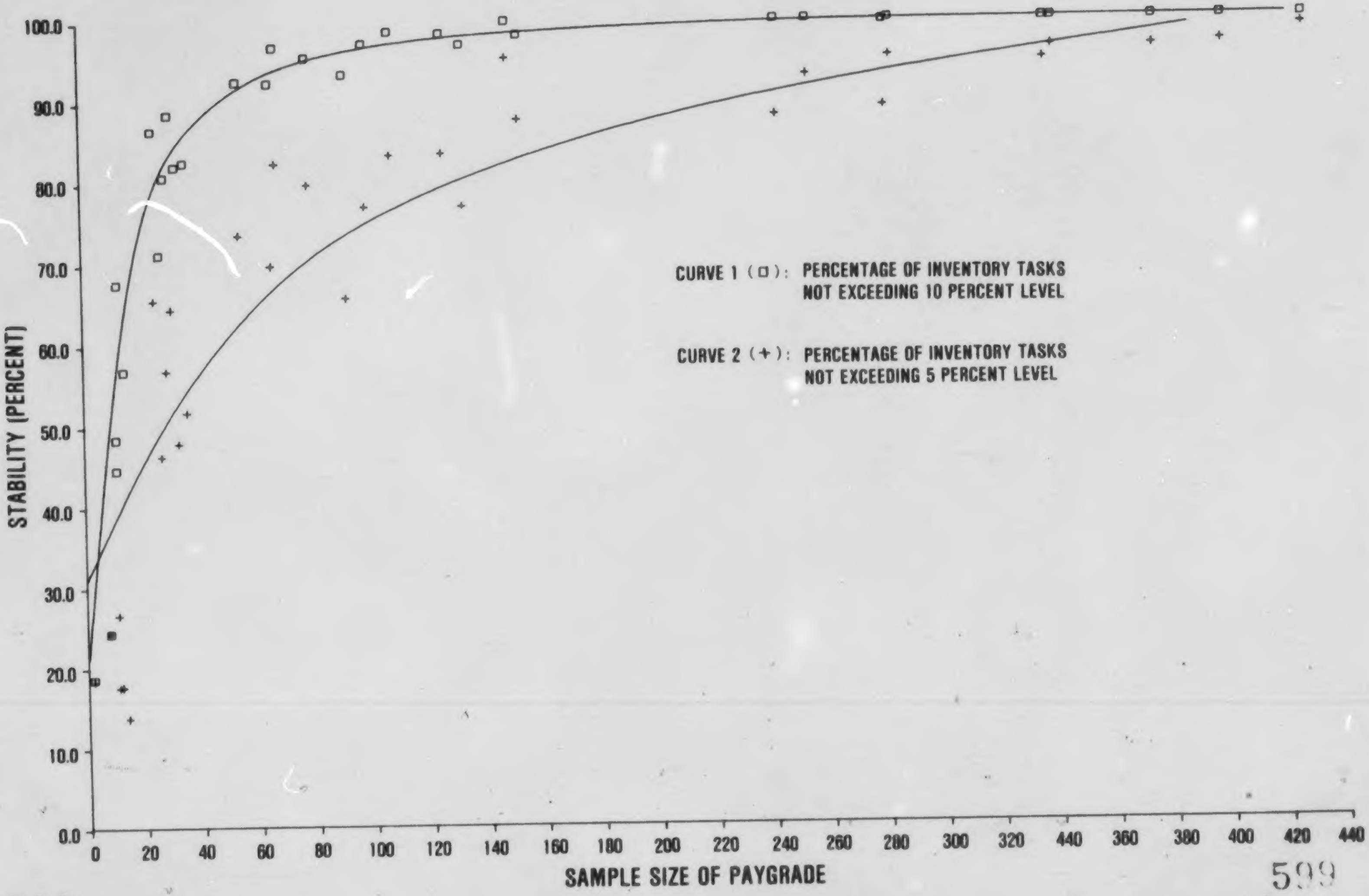
DISPLAY 8



597

596

DISPLAY 9



100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
STABILITY (PERCENT)
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
440
360
380
400
420
440
SAMPLE SIZE OF PAYGRADE
CURVE 1 (□): PERCENTAGE OF INVENTORY TASKS NOT EXCEEDING 10 PERCENT LEVEL
CURVE 2 (+): PERCENTAGE OF INVENTORY TASKS NOT EXCEEDING 5 PERCENT LEVEL

DISPLAY 10
SAMPLE SIZE EFFECT ON STABILITY OF PERCENTAGES OF MEMBERS PERFORMING INVENTORY TASKS

	ACTUAL VALUE STABILITY		RELATIVE VALUE (RANK ORDER) STABILITY
PAYGRADE SAMPLE SIZE	5 PERCENT LEVEL	10 PERCENT LEVEL	CORRELATION
40	58%	87%	.90
100	75%	97%	.97
240	91%	100%	.99
340	96%	100%	.99
440	99%	100%	.99

DISPLAY 11
CONCLUSIONS

---

1. THE TIME-SPENT SCALE CAN BE DELETED TO REDUCE INVENTORY ADMINISTRATION TIME

2. RESPONSES TO CURRENTLY ADMINISTERED INVENTORY SCALES COULD BE USED TO CALCULATE PERCENT PERFORMING DATA

3. THE STUDY'S EMPIRICALLY DEVELOPED GUIDELINES CAN BE USED AS AN AID TO DETERMINE OPTIMAL SAMPLE SIZE

BENCHMARK SCALES FOR COLLECTING TASK TRAINING FACTOR DATA

By
David C. Thomson
and
Kenneth Goody
Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, Texas

Introduction

The Occupation and Manpower Research Division of the Air Force Human Resources Laboratory (AFHRL) is engaged in research into an advanced methodology for determining task training priorities (Christal, 1970; Mead, 1976). One element of this research is the development of benchmark scales for measuring task factors that contribute to training priority decisions. The type of benchmark scale employed is a 9-point scale on which each level is represented by three typical tasks, drawn from a large number of specialties, that illustrate that level. Scales have been developed for three task factors. They are: Probable Consequences of Inadequate Performance, Task Delay Tolerance, and Task Difficulty. In all, three series of scales have been developed, one for specialties with an Administrative or a General (A/G) aptitude requirement, the second for specialties with an Electronic (E) aptitude requirement, and the third associated with a Mechanical (M) aptitude requirement.

At Annex A is an example of one of the nine scales developed and validated over the last two years. The development phase of such a scale has been fully documented and reported by Goody (Psychology in the USAF Symposium Apr 76), Goody and Watson (MTA in Oct 76) and Goody (AFHRL Technical Report 76-15). This paper will not repeat the description of the development phase of the scales, but will address the field testing of the scales, their use and future research areas.

Background

The benchmark scales were conceived as a means to permit measurement of task factors against common frames of reference for various specialties. It was envisaged that a limited number of regression equations using benchmark scale task factor data could be computed, each applying across a number of specialties, that would predict task training priorities. Task factor scales to date have been of a relative nature, in that the ratings given on a task were dependent on the nature of the other tasks in the specialty. While such ratings can be used to predict task training priority within a specialty, a new regression equation must be computed for each specialty.

Field Testing of Benchmark Scales

The benchmark scales, as developed, were field tested by comparing the relative scales rating data with the benchmark scales rating data over at least two specialties for each of the nine scales. The extensive range of the study is tabulated in Table 1, which shows the specialties and sample sizes used in the testing.

Supervisors, randomly drawn from each of the specialties listed, were asked to rate their own career ladder inventories on a single task factor using either the relevant benchmark scale or the relative scale. Using standard techniques, raters were deleted if their task means were significantly ($p \leq .01$) divergent from the sample task means, this fixed selection rule being applied to each rater and each sample. Comparison of benchmark and relative sample sizes and percentage deletions of divergent raters could now be made and a conclusion drawn about the relative efficiency of the scales. These results are shown in Table 2.

The next step in the analysis of the data was to standardize the interrater reliability coefficients so as to suppress the effect of rater response set and permit direct comparisons of the reliabilities. The significance test used was that developed by Haggard (1958). The test requires conversion of the reliability coefficients into Z scores and then a significance test on the difference between the relevant benchmark and relative scale Zs. Results of those tests are tabulated in Table 3.

Finally to test whether raters using the benchmark scales converge on the same vector as they do using the relative scales, the benchmark raw vectors of task means were correlated with the corresponding relative scale raw vectors of task means. Pearson correlation coefficients are tabulated in Table 4.

Findings

Although raters using the benchmark scales have to use technical knowledge outside their past and current job experiences, it was found that on the average only 10% of those raters had to be deleted compared with an average 16% of each sample of raters using the relative scales. This significant difference in percentage rater deletions implies that by using benchmark scales, generally smaller samples can be used, with the associated cost savings benefits, to achieve equally good reliabilities.

At a probability of 0.05, the benchmark rater agreement coefficients are significantly higher than the relative rater agreement coefficients in 14 comparisons, not significantly different in 10 comparisons, and significantly lower in 3 comparisons. Investigation of these later 3 cases showed that the raters were not sufficiently familiar with the tasks on particular benchmark scales to be able to make reliable ratings. Future research needs to address this question as to which subsets of raters are sufficiently knowledgeable to be able to reliably use the benchmark scales.

603

Of the 27 Pearson correlation coefficients, only nine are below .85 and of these only two are below .72. In those two cases, low relative scale interrater reliabilities contributed strongly to the poor correlations. But as high correlations were generally found to be the order, it can be concluded that raters using the benchmark scales do rank the tasks in the same order as raters using the traditional relative scales.

Future Research

Although not discussed in depth in this paper, the problem of raters not being sufficiently familiar with the tasks listed on the benchmark scales does exist. Future research needs to address this problem. One way around the problem is to accept the technology for what it is, and develop benchmark scales to address questions across specialties in a limited number of similar specialties (e.g. aircraft systems maintenance) such as exist in a career field.

There are some indications in the research data that supervisors, using the benchmark scales and rating their own career ladder inventory tend to inflate their ratings. That is they tend to indicate that the task difficulty is higher than it really is, that the acceptable delay before a task must be performed is smaller than it really is, and that the consequences of not doing a task properly are much worse than they really are. Furthermore, this inflation does not appear to be constant or even predictable. Research must address and solve this problem before task factor comparisons across large numbers of specialties can be made. Developing special benchmark scales for use within career fields may help, since the problems of rater inflation and raters being unfamiliar with tasks on the scales should be less.

Conclusion

Benchmark scales will allow experienced raters to provide better interrater agreement than do the relative scales and the desired level of stability of the means is obtained more efficiently as fewer rater deletions are necessary. Furthermore, these same raters preserve the correct rank ordering of the tasks on the different task factors. However, considerable effort and care is needed to make certain that the intended raters have a reasonable amount of familiarity with the tasks that define the various points on each benchmark scale. Some inflation of ratings for the raters' own specialty should be expected.

REFERENCES

Christal, R. E. Implications of Air Force occupational research for curriculum design. In B. B. Smith & J. Moss, Jr. (Eds.), *Report of a Seminar: Process and techniques of vocational curriculum development*. Minnesota Research Coordinating Unit for Vocational Education, University of Minnesota, Minneapolis, MN., April 1970, 27-61.

Goody, K. Task factor benchmark scales for use in determining task training priority. Paper presented in 5th Psychology in the AF Symposium, U. S. Air Force Academy, CO., 8-10 April 1976.

Goody, K. and Watson, W. J. Task factor benchmark scales for use in determining training priority. Paper presented at the 17th Annual Conference of the Military Testing Association, U. S. Army, Indianapolis, IN., 15-19 September 1975.

Goody, K. *Task factor benchmark scales for training priority analysis: Overview and developmental phase for administrative/general aptitude area*. AFHRL-TR-76-15. AD-A025 847. Brooks AFB, TX.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976.

Haggard, E. A. *Intraclass correlation and the analyses of variance*. New York: Dryden Press Inc., 1958, 25-26.

Mead, D. F. Establishing job task training priorities. Paper presented at the First International Learning Technology Congress, Washington, D.C., 20-24 July 1976.

Table 1

NUMBER OF RATERS BY FACTOR AND TYPE OF RATING SCALE
FOR 11 AFS IN FINAL VALIDATION STUDY

Air Force Specialty	Minimum Aptitude Requirement	Number of Raters					
		Consequences		Delay Tolerance		Task Difficulty	
		Bench.	Relative	Bench.	Relative	Bench.	Relative
293X3 Radio Operator	A60	51	45	49	50	49	78
651X0 Procurement	A70	67	61	71	63	59	101
531X5 Non Destructive Inspection	G50	61	-	67	-	67	55
906X0 Medical Administration	G60	77	105	87	104	101	78
304X4 Ground Radio Communication Equip	E80	66	60	57	58	55	122
304X0 Radio Relay Equip	E80	39	35	39	50	44	89
423X4 Pneudraulic Repair	E, M40	60	-	71	-	69	[illegible]
552X5 Plumbers	M40	69	82	66	62	69	116
423X1 Environmental Systems	M40	52	33	52	34	52	77
427X5 Airframe Repair	M40	71	63	77	65	74	75
631X0 Fuel Specialtists	G, M40	71	-	71	-	74	75
	Average	62	61	64	61	65	85

600

Table 2

PERCENTAGE OF RATERS DELETED

Specialty/Aptitude	Consequences		Delay Tolerance		Task Difficulty	
	Benchmark	Relative	Benchmark	Relative	Benchmark	Relative
293X3 Radio Operator A60	8	20	12	24	16	24
651X0 Procurement A70	12	5	17	19	5	22
906X0 Medical Administration G60	3	5	5	5	2	6
531X5 Non Destructive Inspection G50	7	-	25	-	6	16
304X4 Ground Radio Communication Equipment E80	5	7	14	16	0	11
304X0 Radio Relay Equipment E80	5	3	0	20	2	11
423X4 Pneudraulic Repair E, M40	10	-	11	-	10	11
552X5 Plumbers M40	16	6	28	27	3	34
423X1 Environmental Systems M40	6	6	10	29	6	14
427X5 Airframe Repair M40	20	5	21	20	11	28
631X0 Fuel Specialists G, M40	20	-	11	-	7	25
Average Percentage Raters Deleted	10	7	14	20	6	18
Average Final Sample Size	52	54	51	46	56	53

607

608

Table 3

COMPARISON OF STANDARDIZED R_{11} VALUES DERIVED FROM BENCHMARK AND RELATIVE SCALE DATA

Specialty/Task Factor	Benchmark Standard. R_{11}	Relative Standard. R_{11}	Benchmark F_{11}	Relative F_{11}	Bench. K	Relat. K	Number of Tasks	$\frac{Z_B - Z_R}{\sigma_{Z_{B-R}}}$	Probabili
TASK DIFFICULTY									
293X3 Radio Operator	.357	.221	16.6	11.9	28.0	38.5	345	3.02	<.05
651X0 Procurement	.395	.321	36.3	30.2	54.1	61.7	328	1.65	.10
531X5 Non Destructive Inspec.	.342	.185	30.8	10.5	57.3	42.0	230	8.03	<.05
906X0 Medical Administration	.460	.428	80.3	35.2	93.1	45.8	813	11.67	<.05
304X4 Ground Radio Comm. Equip.	.412	.335	35.5	32.9	49.3	63.3	730	1.03	.30
304X0 Radio Relay Equipment	.394	.259	26.9	22.7	39.7	62.1	322	1.49	.14
423X4 Pneudraulic Repair	.292	.297	23.9	23.3	55.4	52.8	575	.28	.78
552X5 Plumbers	.333	.310	33.1	34.1	64.3	73.5	407	- .30	.76
423X1 Environmental Systems	.280	.312	17.9	23.5	43.4	49.7	736	-3.69	<.05
427X5 Airframe Repair	.357	.302	36.1	22.1	63.1	48.7	252	3.85	<.05
631X0 Fuel Specialists	.345	.340	35.7	22.5	65.7	41.8	374	4.40	<.05
CONSEQUENCES OF INADEQUATE PERFORMANCE									
293X3 Radio Operator	.369	.314	20.8	16.1	34.0	33.0	345	2.34	<.05
651X0 Procurement	.222	.217	16.4	16.7	53.9	56.6	328	- .18	.86
906X0 Medical Administration	.230	.258	22.5	34.2	72.0	95.4	813	-5.91	<.05
304X4 Ground Radio Comm. Equip.	.281	.265	23.6	20.7	57.7	54.6	730	1.75	.80
304X0 Radio Relay Equipment	.403	.247	25.4	11.7	36.3	32.8	322	6.82	<.05
552X5 Plumbers	.159	.265	11.7	27.3	56.9	72.8	407	-8.43	<.05
423X1 Environmental Systems	.305	.277	20.5	11.5	44.4	27.4	736	7.74	<.05
427X5 Airframe Repair	.382	.274	35.4	23.4	55.6	59.3	252	3.24	<.05
TASK DELAY TOLERANCE									
293X3 Radio Operator	.370	.325	19.9	15.5	32.2	30.0	345	2.27	<.05
651X0 Procurement	.276	.246	22.4	17.0	56.1	48.8	328	2.49	<.05
906X0 Medical Administration	.258	.251	28.9	30.8	80.2	89.0	813	- .91	.37
304X4 Ground Radio Comm. Equip.	.282	.129	19.1	7.7	46.0	45.5	730	12.09	<.05
304X0 Radio Relay Equipment	.283	.268	15.3	14.4	36.4	36.6	322	.56	.57
552X5 Plumbers	.168	.155	10.9	9.0	49.1	43.5	407	1.92	.05
423X1 Environmental Systems	.243	.217	14.5	6.8	42.1	20.9	736	10.07	<.05
427X5 Airframe Repair	.330	.246	29.3	17.7	57.5	51.3	252	3.93	<.05

Table 4

PEARSON CORRELATION COEFFICIENTS

Specialty/Aptitude	Consequences	Delay Tolerance	Task Difficulty
293X3 Radio Operator A60	.91	.92	.59
651X0 Procurement A70	.89	.91	.93
906X0 Medical Administration G60	.85	.94	.94
531X5 Non Destructive Inspection G50	-	-	.73
304X4 Ground Radio Communications Equipment E80	.92	.73	.92
304X0 Radio Relay Equipment E80	.87	.90	.78
423X4 Pneudraulic Repair E, M40	-	-	.82
552X5 Plumbers M40	.82	.47	.89
423X1 Environmental Systems M40	.94	.72	.92
427X5 Airframe Repair M40	.81	.85	.93
631X0 Fuel Specialists G, M40	-	-	.91

ANNEX A

TASK DELAY TOLERANCE

(Electronic)

DEFINITION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it.

BENCHMARK SCALE

Level 9 – Most Tolerance of Delay – Do when ready

Clean or paint missile facilities or equipment (Missile Systems Maintenance Specialist)

Wash, clean or inspect maintenance vehicles (Flight Facilities Equipment Specialist)

Write test questions (Avionic Inertial and Radar Navigation Systems Specialist)

Level 8

Revise technical orders or indices (Weather Equipment Repairman)

Inventory bench stock, equipment or supplies (Flight Facilities Equipment Specialist)

Maintain electrical storage battery records (Telephone Switching Equipment Repairman)

Level 7

Clean parts or components using solvents (Avionic Navigation Systems Specialist)

Locate part or stock numbers in federal supply catalogs (Precision Measurement Equipment Laboratory Specialist)

Prepare or maintain Explosive Ordinance Disposal reports (Munitions Disposal Specialist)

Level 6

Change oil in antenna drive assemblies (Air Traffic Control Radar Repairman)

Analyze computer logic diagrams (Electronic Computer Systems Repairman)

Trace underground power cables using cable test set (Electrical Power Line Specialist)

Level 5

Tighten bolts or nuts to specified torques (Missile Systems Analyst Specialist)

Troubleshoot aircraft radio switching systems (Avionic Communications Specialist)

Perform operational tests on angle-of-attack or side-slip transmitters (Integrated Avionics Component Specialist)

Level 4

Test or check safety devices such as valves, regulators, or alarms on biomedical equipment (Biomedical Equipment Maintenance Repairman)

Load nuclear bombs, warheads or reentry vehicles onto transport aircraft (Nuclear Weapons Specialist)

Repair or adjust aircraft cockpit latches or locks (Aircrew Egress Systems Repairman)

Level 3

Perform inflight analysis of malfunctions in automatic tracking radar (Auto Tracking Radar Repairman)

Target or retarget guided missiles (Missile Systems Analyst Specialist)

Install nuclear weapon fusing systems (Weapons Mechanic)

Level 2

Perform nuclear bomb safety checks (Nuclear Weapons Specialist)

Monitor aircraft engine instruments during flight (Flight Engineer Specialist)

Check aircraft for armament safety (Weapons Control Systems Mechanic)

Level 1 – Least Tolerance of Delay – Must do immediately

Conduct emergency shutdown of missile launch facility (Missile Systems Analyst Specialist)

Render aircraft emergency egress systems safe after crash (Aircrew Egress Systems Repairman)

Perform emergency shutdowns of high pressure boilers (Plant Operator)

SECTION 7

PERSONNEL SELECTION

WEIGHTED SELECTION SYSTEM FOR AFROTC APPLICANTS -- PERSPECTIVE AFTER SECOND YEAR OF USE

Lieutanant Colonel David K. Jackson
Mr. M. Meriwether Gordon, Jr.

At last year's meeting of the Military Testing Association, AFROTC representatives reported on the "Development of a Weighted Selection System" for admitting applicants into the AFROTC Professional Officer Course. This course is the last two years of the four-year AFROTC Program and leads to an Air Force commission on graduation.

The weighted system has come to be known as WPSS (pronounced WEEP-us) standing for Weighted POC Selection System. The system was developed on the basis of the findings of a model selection board held at Maxwell AFB. With the assistance of Human Resources Laboratory, statistical "policy capturing" techniques were applied to the board's findings. The variables considered by the board in rank-ordering applicants for the program were processed through a system known as "Hierarchical Grouping" and assigned weights in accordance with the contribution each made to the individual's rank-order. About ninety variables were considered and reduced to eleven which were identified as contributing significantly. Those variables and their weights together with the number of points each contributes to the total Quality Index Score (QIS) derived by the system are listed in Table 1. Note that the computations are based on an assumed mean of each of the variables. These are the actual means that were attained on each after all applications were received and the overall means computed.

TABLE 1
Eleven Variables
Constituting the QIS

VARIABLE	MEAN	WT	POINTS	% SCORE
AFOQT--Quality Composite	45.5	0.1381	6.28	(8.4)
SAT Score	1054.2	0.0245	25.83	(34.5)
Cumulative GPA	277.3	0.1005	27.87	(37.2)
PAS Rating	3.2	1.7975	5.75	(7.7)
ASTIN Rating	3.5	0.7172	2.51	(3.3)
AFROTC GPA	224.2	0.0130	2.91	(3.9)
AFOQT--Quantitative	47.0	0.0459	2.16	(2.9)
Type Program	0.6	1.5837	0.95	(1.3)
Academic Major	0.4	2.5949	1.04	(1.4)
Number of Applicants	36.6	0.0222	0.81	(1.1)
Applicants Rank	14.7	-0.0870	-1.28	(-1.7)
	Quality Index Score:		74.83	

The Officer Quality Composite of the Air Force Officer Qualifying Test is shown as contributing 8.4% to the total score while the SAT is shown as contributing 34.5%. These figures merit additional qualification. Where students have ACT scores instead of SAT scores, the scores are converted to SAT equivalents. If the students have both ACT and SAT scores, they may convert the ACT scores to SAT equivalents if the conversion results in a higher score. Where students lack either ACT or SAT scores, they are allowed to convert their Officer Quality scores if they benefit thereby. Indeed, they may convert the Officer Quality Score in any case where this would be advantageous to them. In addition, the AFOQT Quantitative score--a sub-test of the Officer Quality score--is counted separately and contributes 2.9%. Thus, the Officer Quality Score may contribute much more heavily than the figures seem to indicate. Standardized tests in toto--ACT/SAT/AFOQT--contribute 45.8% of the total Quality Index score with the cumulative GPA as the next highest contributer at 37.2%. Standardized test scores and the grade point average in combination contribute 83% of the total Quality Index Score.

The Professor of Aerospace Studies (PAS) rating is done on a scale of 0 to 4 and amounts on the average to 7.7% of the total score. The PAS can exert a little additional influence on the overall score through the rank-order of the applicant. The rank-ordering is done either by the PAS or by a local board of which the PAS is usually a member. (Note the negative weight of the applicant's rank among those ranked).

The "Astin Rating" is a college selectivity rating on a scale of 1 to 7 devised and published by Dr. Alexander W. Astin in his book entitled Predicting Academic Performance in College.

The "Type Program" variable provides the applicant some credit for participation in the four-year programs over the two-year program. Its value is either 0 or 1 times its weight.

The "Academic Major" variable provides credit to those applicants with desired scientific/engineering academic majors. Again, its value is either 0 or 1 times its weight.

The Pilot and Navigator-Technical Composites of the AFOQT are not factors in the QIS. Nevertheless, they are powerful as qualifiers for the program since applicants are not eligible for consideration under the WPSS as potential pilots or navigators unless they have attained at least the minimum requirements set by the Air Force on these composites.

Since standardized test scores in combination are the most heavily weighted factor in the system and no statistical distinction is made between SAT, ACT, and Officer Quality Scores, the correlation matrix in Table 2 is of interest. The matrix also includes the Verbal and Quantitative sub-composites of the Officer Quality score, the applicant's grade-point-average (4.00 scale), and the Quality Index Score (QIS) derived by the system. Only those applicants possessing both SAT and

ACT scores were used. The coefficients to the left were derived from 287 applicants possessing all three scores who applied in FY 77. The coefficients to the right (in parentheses) were derived from 341 such applicants in FY 78.

TABLE 2
Correlation Matrix
(Pearson Product-Moment)

	ACT	SAT	OQC	VERB	QUANT	GPA	QIS
ACT	—	.85(.83)	.73(.67)	.66(.68)	.58(.57)	.21(.26)	.74(.72)
SAT	.85(.83)	—	.72(.66)	.65(.66)	.58(.58)	.27(.29)	.76(.73)
*OQC	.73(.67)	.72(.66)	—	.74(.69)	.77(.77)	.21(.26)	.81(.79)
VERB	.66(.68)	.65(.68)	.74(.70)	—	.45(.42)	.21(.17)	.62(.56)
QUANT	.58(.57)	.58(.58)	.77(.77)	.44(.42)	—	.15(.24)	.67(.73)
GPA	.21(.26)	.27(.29)	.21(.26)	.21(.17)	.15(.24)	—	.59(.63)
QIS	.74(.72)	.76(.73)	.81(.79)	.62(.56)	.67(.73)	.59(.63)	—

It is interesting to note that the ACT and AFOQC predict the grade-point-average to an equal degree while the SAT--probably the most highly and systematically standardized test in existence--predicts it only slightly better. Indeed, one might reasonably contend that the three tests are about equally predictive of academic success as measured by the grade-point-average.

Some individuals have expressed surprise and dismay at the seemingly low correlations between standardized test scores (SAT/ACT/OQC) and the grade-point-average. All of these tests purport to predict academic success.

It must be remembered that among students taking the SAT and ACT many low scorers are dissuaded from going to college and are not present to be included in the validity data. Many others for whom the tests accurately predicted failure did not survive in college long enough to be included in this group of applicants for the advanced AFROTC program. Finally, any students for whom the test inaccurately predicted failure are still present and count against the test's validity. For those reasons, at this stage of the game--the end of the sophomore year--these coefficients should be considered quite good.

---

*The OQC scoring scale is restricted in range in comparison with the ACT an SAT scales. If OQC scores were free to vary over the same range as their A and SAT counterparts, their correlation with the other two tests (and with the GPA) would probably be higher.

It is also interesting to note that the Air Force Officer Quality score stands up well in the company of its highly prestigeous competitors. Indeed, any one of the three scores could readily be accepted as a predictor of academic success in lieu of either of the other two.

* *

For many years, AFROTC allowed its separate detachments to make their selection for entry in the Professional Officer Course (POC) locally in the same manner that other college programs admit their applicants. Local selection was allowed despite the knowledge that the lowest individual Officer Quality score at some of the highly selective institutions was higher than the highest score at some of the low selectivity institutions. However, the decline in the number and quality of applicants that ensued after the advent of the all-volunteer force made it apparent that AFROTC would have to exert strong quality controls to insure a continued [redacted] officer corps. Central selection of applicants seemed a necessary measure though one that AFROTC was reluctant to take.

The new WPSS has proved to be a more than adequate compromise between local and central selection. Under the new system, Air Force Professors of Aerospace Studies (Detachment Cammanders) are allowed to fill their enrollment quotas locally with students possessing Quality Index Scores of sixty-three or above prior to a given cut-off date. The names of those with scores above sixty-three in excess of quotas or who apply after the cut-off date and those with scores below sixty-three are submitted to Maxwell AFB for central selection. The names of those selected locally are also submitted for official confirmation. Thus, all or nearly all, of the selections are still made locally at the more selective institutions, while selections at the less selective institutions are partially made by central board. While in theory a local selectee might be thrown out in favor of a more highly qualified central selectee, this in fact did not happen.

The new system allows AFROTC to enjoy simultaneously the best aspects of both local and central selection. Indeed, it is possible to make the seemingly paradoxical assertion that while all the selections are made centrally, most are still made locally. That is to say that all the selections made locally are those that would have been made by central selection and do not be come final until confirmed by the central board.

* *

One of the problems confronting the central selection boards has been the high incidence of drop-outs among applicants already selected. About 22% of selectees did not subsequently enroll, this has kept the central selection boards in action throughout the summer months and up to the starting day of class and even beyond. A great deal of conjecture occurred about why this should be so. One hypothesis was that the drop-outs were occurring among the higher quality applicants who had wider and better alternatives than their less talented fellows and were being distracted by offers from competitors. As reasonable as this hypothesis sounded, it has proved to be largely untrue. Drop-outs are about equal in quality to those who remain as may be seen from the data in Table 3.

TABLE 3
Comparison of Applicants
(Applied, Selected, and Selected/Dropped)

	No.	OQC	GPA	*SAT-EQ	QIS
a. Applied	4613	50.6	2.79	1090	77
b. Selected	4470	52.0	2.81	1098	78
c. Selected Dropped	1234	50.0	2.83	1089	76

However, the term "dropped," as employed here includes all selectees who for some reason after selection failed to enter the program when classes began. The failure may have been totally involuntary as would be the case with academic eliminations from the institution, medical disqualifications, or headquarters disapproval of a request to waive some disqualifying characteristic. (Arrest, drug-abuse, etc.). Some drop-outs might be called semi-voluntary such as students who could not gain entry in the category desired (pilot or navigator) or who failed to receive an anticipated scholarship. Other drop-outs are entirely voluntary such as those who simply lose interest, change their minds about enrolling, or who enroll in an Army or Navy program.

Reasons for drop-out insofar as they could be determined and the various quality measures associated with each are as detailed below:

TABLE 4
(Reason for Drop-Out)

	No.	%	OQC	GPA	SAT-EQ	QIS
a. Academic	100	8.1%	51	2.43	1086	73
b. Physical	159	12.9%	49	2.73	1070	73
c. Quota Competition	30	2.4%	38	2.69	1023	70
d. HQ Disqualified	25	2.0%	46	2.65	1088	74
e. Outside Competition	216	17.5%	54	2.93	1097	78
f. Field Tng Elim	87	7.1%	51	2.85	1089	75
g. Personal	582	47.2%	50	2.91	1093	76
h. Scholarship	11	.9%	55	2.72	1120	78
i. Other + Unknown	24	1.9%	52	2.68	1105	76
j. Overall	1234	100%	50	2.83	1089	76

---

*SAT-EQ includes actual SAT scores or ACT/OQC conversions to SAT. The mean is computed without distinction between actual and converted scores.

The two largest groups of drop-outs are those citing "Outside Competition" and those citing "Personal" as the reason for drop-out. These are also the two groups among which the reason for drop-out is strictly voluntary. Therefore, these drop-outs merit more detailed examination. Their characteristics are outlined in Tables 5 and 6.

TABLE 5
(Drop-Outs--Outside Competition)

Received a better offer from:

	No.	%	OQC	GPA	SAT-EQ	QIS
a. Civilian Source	140	64.8%	54	2.96	1103	78
b. Other Military	54	25.0%	56	2.85	1097	77
c. Other Government Agency	8	3.7%	39	2.94	1034	74
d. Unknown/Other	14	6.5%	48	3.00	1070	76
Total	216	100%	54	2.93	1097	78

TABLE 6
(Drop-Outs--Personal Reasons)

	No.	%	OQC	GPA	SAT-EQ	QIS
a. Lost Interest	360	61.8%	50	2.91	1098	77
b. Peer Pressure	4	.7%	35	3.23	1110	79
c. Family Problems	62	10.7%	43	2.90	1049	74
d. Financial Problems	41	7.0%	45	2.88	1066	73
e. Active Duty (not Released)	9	1.5%	47	2.73	1059	75
f. Girl/Boy Friend	23	4.0%	42	2.89	1083	75
g. Religion	7	1.2%	58	2.88	1124	81
h. Unknown	29	5.0%	60	2.85	1118	77
i. Other	47	8.1%	58	2.94	1126	80
Total	582	100%	50	2.91	1093	76

The figures in Table 5 by no means define AFROTC's problems with competition from outside agencies. The loss of 140 prime selectees to civilian competitors is regrettable. What we do not know is how many were lost before they ever applied for selection.

What does become apparent under analysis of the data is that 360 high quality selectees dropped-out simply because they "lost interest" between the time of selection and the first day of class. If the individual detachments, by a vigorous follow-up campaign, could succeed in reducing this figure by half, they might succeed in reducing the over all drop-rate from 28% to 24%. Getting the rate much lower than 24% would not seem a realistic goal.

Characteristics of the FY 78 WPSS selectees (for FY 80 graduation) are as displayed in Tables 7 and 8. Table 7 shows the mean scores by sex and race of fall 78 enrollees on the Officer Quality Composite and Quantitative Composite of the Air Force Officer Qualifying Test and their mean grade-point-averages on the four-point scale. Table 8 shows the same information by enrollment category: pilot, navigator, missile specialist, science-technical, and other.

TABLE 7

(Mean Standardized Test Scores and Grades of WPSS Selectees for FY 78, by Sex and Race)

	Overall	Male	Female	Caucasian	Black	Other
Total "N"	3211	2691	520	2770	330	111
SAT	1099	1105	1065	1119	949	1032
OQC	52	54	44	56	26	36
QUANT	55	57	46	58	33	46
GPA	2.80	2.77	2.91	2.81	2.69	2.76

TABLE 8

(Mean Standardized Test Scores and Grades of FY 78 WPSS Selectees by Category)

	Overall	Pilot	Navigator	Missile	Tech/ Science	Non-Tech
Total "N"	3211	848	285	298	902	878
SAT	1099	1141	1092	1064	1152	1017
OQC	52	61	50	47	61	37
QUANT	55	65	56	48	68	36
GPA	2.80	2.84	2.58	2.68	2.90	2.76

Table 9 shows the improvement in Officer Quality Score since the pre-WPSS days in fiscal year '67.

TABLE 9

	OQC	SAT
a. FY 76 (Pre-WPSS)	41	-
b. FY 77 (1st Yr WPSS)	49	1087
c. FY 78 (2nd Yr WPSS)	52	1099

AFROTC is justified in concluding that the system has helped identify quality applicants and allowed AFROTC to select the highest quality applicants. AFROTC intends to continue to use, study, and refine the system for the foreseeable future.

THE DEFENSE LANGUAGE APTITUDE BATTERY

Robert G. Henderson
Education Specialist

A paper presented to the
Military Testing Association
October 1978

DEFENSE LANGUAGE INSTITUTE
FOREIGN LANGUAGE CENTER

Presidio of Monterey, CA 93940

ABSTRACT

The Defense Language Aptitude Battery (DLAB) was introduced in 1977 for use by the Defense Language Institute Foreign Language Center to screen potential candidates for training in over thirty foreign languages. Its predictive validity for success in foreign language training is higher than its predecessor test and two commercially-available language aptitude tests. Differential prediction by language was studied as a part of the validation research. That hypothesis, when using DLAB, was not sustained.

THE DEFENSE LANGUAGE APTITUDE BATTERY

The Defense Language Institute Foreign Language Center is located at the Presidio of Monterey, California, and operates under a direct charter from the Department of Defense which names the Department of the Army as Executive Agent for operation of the school.

At Monterey some thirty foreign languages (expandable to fifty languages) are offered to a group of some 2,200 students at any given time. We graduate between three and four thousand students annually. Our students are predominantly officer and enlisted personnel from the four military branches plus a smattering of civilian students from other federal agencies. Spouses of students are also invited to attend class on a space-available basis.

Like most military schools, the Institute is concerned with cost-effective operations while producing the best-qualified linguists possible. One method employed is to attempt to predict, and therefore control, student attrition for academic reasons. The Defense Language Aptitude Battery (DLAB) is used for this purpose.

Each military branch has its own recruiting criteria for physical and mental standards. Whether the individual is a first-term service man or woman, or someone with a number of years of military service, candidates for foreign language training at Monterey are required to take the DLAB. General and flag officers are excused from this requirement.

Exercising his authority over technical control of the Defense Foreign Language Program, the Commandant, DLIFLC, sets the minimum scoring criteria (cutting score) on DLAB that represents eligibility for training. Waivers may be granted at the discretion of the Commandant. The test is usually administered at Armed Forces Entrance and Examination Stations, or at Lackland Air Force Base, Texas. Some testing is done at Monterey.

The views of the author do not purport to reflect the official position of the Department of Defense, the United States Army, or the Defense Foreign Language Center.

The Defense Language Aptitude Battery is the successor to the Defense Language Aptitude Test (DLAT), which was used for over twenty years. DLAB was implemented in the summer of 1977 and the use of DLAT rescinded at that time. The two tests differ in several significant ways.

While both tests are paper and pencil tests using a multiple-choice format, DLAB also contains an audio component. DLAT did not. This was incorporated into the test design because of the teaching methodology used at Monterey. This is predominantly the audio-lingual method, which places a considerable burden on listening, as opposed to cognitive-code, grammar-translation and other traditional foreign language teaching methodologies.

The old DLAT was prepared in two alternate forms. The construction and equating of alternate test forms is an expensive and time-consuming operation. The purpose of constructing alternate forms is to mitigate the problems of compromise and practice effect when personnel are retested. Experience on DLAT indicated that few individuals ever requested a second test administration. Further, the unique design of DLAB is such that compromise short of possessing the answer key would be difficult. As an additional safeguard, test length was extended from fifty-nine items on DLAT, to 119 items on DLAB.

DLAT required about thirty minutes to administer and DLAB requires about ninety minutes. This has caused some difficulty at the AFEES, where each processing minute is very important. We are now conducting item analysis on a sample of approximately 2,000 answer sheets to investigate the possibility of reducing test length without disturbing validity or reliability.

DLAB enjoys one great advantage over DLAT. That is, the meticulous field validation of the test, its revision, and subsequent cross-validation using external criteria on a large population of our students. The result is that DLAB has a correlation with student achievement of .50, as opposed to .35 for DLAT.

The size of the population taking the test at AFEES and military bases worldwide is known to us, but fluctuations in the size of the population are dependent upon variables not under our control. With known "pass rates" associated with a given cutting score we can establish a fair idea of the student eligibility pool for that cutting score.

An abrupt decrease in the population passing the test without a corresponding decrease in linguist requirements would suggest that the cutting score be lowered. While this would increase the relative eligibility pool it would also be accompanied by a rise in academic attrition. Thus, we attempt to peg the cutting score at a point that will permit the appropriate number of individuals to become eligible while maintaining the lowest possible predicted academic attrition rate.

Periodically, DLAB scores are compared with classroom performance by our students using the final course grade as the criterion. Combined with input numbers and attrition data, we can then establish an optimum cutting score. Based upon simulated prediction, the current raw cutting score of sixty produces an eligibility pool of approximately twenty per cent of those actually tested. In actual performance the eligibility yield has been slightly higher, between twenty three and twenty four per cent. We suspect that this small fluctuation is due to changes in the total test population. The groups upon whom the test was normed back in the early nineteen seventies was almost exclusively male and predominantly white. In recent years the recruit population has included growing numbers of females and blacks. Earlier this year, at the request of the Defense Department, we performed a preliminary study on the limited number of answer sheets then on hand to determine if there were significant differences in the test population in terms of military branch, ethnic origin and gender. That information did indicate differences. In general, DLAB scores were slightly higher for Navy personnel, whites and females. The first full cycle of these individuals are now completing their foreign language training. We are keenly aware of the social, legal and cost-effective operational implications of these preliminary findings. When adequate numbers of personnel have completed the training cycle we will be obligated to investigate these variables further.

As a selection, classification and screening device, DLAB does not operate in a vacuum. There are many other related factors. For instance, the military branches may impose minimum attainment standards on the Armed Services Vocational Aptitude Battery (ASVAB) before individuals are permitted to take DLAB. Auditory acuity is only grossly examined now. We would like to improve the way this is measured. Our students are, by and large, volunteers.

This surely impacts on motivation and attitudes. Many of our graduates must enter fields of work with sensitive security requirements. Therefore, their backgrounds must indicate a high probability of being granted a security clearance before a lot of money is invested in their training. Other than aptitude and learning capacity, these variables are not subject tc the control of DLIFLC.

One other variability interested us when the test was designed. We hoped that the special features of the test might permit some indication of differential aptitude across language families or, perhaps, for individual languages themselves. This proved to be a phantom, probably due to a series of uncontrolled, or even unrecognized, variables. For example, American English is the native language for most of our students. Based upon both lengthy experience and intuition, we generally expect the Romance languages to be easiest for our students to learn, the Slavic group somewhat more difficult, and the Arabic and Oriental languages the most difficult. Unfortunately, there are numerous linguistic differences within these generalities that may enhance or impede the learning process for the native American English speaker. To cite but two examples, our students have considerable problems with tone languages (e.g., Thai, Chinese) and those with unique writing systems (e.g., Arabic, Japanese). And, despite a commitment to rather singular teaching methodology and environment, each course of instruction differs in many ways from the others in areas not related to the language itself.

The initial establishment of the cutting score for DLAB was arbitrary, but not set without considerable information. My, colleague, Mr. Thain, will discuss the techniques employed as well as the rationale for using standard scores for DLAB, instead of raw scores, which were used with DLAT.

For those of you who would like to further examine both the DLAB design concept and the validation procedures, we have a limited number of booklets here at the front table containing the technical reports.

I thank you for your attention and will be happy to respond to any questions you may have.

ΠΤΟΛΕΜΕΙ
DEFENSE LANGUAGE INSTITUTE

DEFENSE LANGUAGE APTITUDE BATTERY (DLAB)

DESCRIPTION

PART I - BIOGRAPHICAL INVENTORY

PART II - RECOGNITION OF STRESS PATTERNS

PART III - FOREIGN LANGUAGE (FL) GRAMMAR

PART IV - FL CONCEPT FORMATION

ACTUAL TEST

Items	119
Practice Items	7
Total Items	126

CUT OFF SCORES FOR ENTRY

RAW	STANDARD
60	89

630

DEFENSE LANGUAGE APTITUDE BATTERY
PREDICTIVE VALIDITIES FOR INDIVIDUAL LANGUAGE COURSES

	ZERO ORDER CORRELATIONS			
		DLAB PREDICTIVE VALIDITY		DLAT PREDICTIVE VALIDITY
LANGUAGE	N	r	N	r
ARABIC	153	.400	140	.210
CHINESE-MANDARIN	85	.624	75	.244
CZECH	86	.640	70	.503
FRENCH	72	.550	43	.311
GERMAN	106	.428	99	.228
KOREAN	92	.547	78	.467
RUSSIAN	86	.678	73	.570
SPANISH	83	.594	66	.507
THAI	51	.521	27	.398
VIETNAMESE	38	.433	37	.040
TOTALS	852	.541	708	.348

COMPUTER SIMULATION OF DLI SELECTION PROBLEM

DLAB CUTTING SCORE	BASE SAMPLE INELIGIBLE %	ATTRITION %	AVERAGE GRADE %
40	42.2	17.8	78.7
42	46.5	16.7	79.1
44	50.7	15.3	79.6
46	54.6	14.0	80.0
48	58.8	12.7	80.5
50	62.9	11.5	81.0
52	66.9	10.5	81.4
54	70.6	10.0	81.9
56	74.0	8.6	82.3
58	77.2	7.7	82.8
60	80.3	7.1	83.3
61	81.7	6.7	83.6
62	83.1	6.1	83.9
63	84.4	5.6	84.2
64	85.5	5.2	84.4
65	86.7	4.8	84.7
66	87.9	4.4	85.0
67	88.9	4.0	85.2
68	90.0	3.9	85.5
69	90.1	3.5	85.9

DEFENSE LANGUAGE INSTITUTE
FOREIGN LANGUAGE CENTER

DEFENSE LANGUAGE APTITUDE BATTERY

Current (April 1978) Statistics

	N	Mean*	Standard Deviation
Total Test Population	24,633	51.6	15.5
Army Sample	870	52.0	17.2
Air Force Sample	1,560	51.3	14.4
Navy Sample	333	55.2	17.8
Marine Sample	(Too Small to Include)		

*Maximum Possible Score = 119

Test Reliability Estimate (N=4,000) .91
(Kuder-Richardson Formula 21)

Passing Rate
(Raw Score Cut-Off = 60)

Test Population	Number of Answer Sheets	Number Pass	Eligible Per Cent
All Services	76.837	14,534	24.9%
Army	43,428	6,686	15.4%
Air Force	33,085	7,744	23.4%
Navy	310	100	32.2%
Marines	14	4	28.6%

Correlations of Predictors With Average Grade
For 1969 - 1971 Sample (879 Cases) in 12 Languages

	Predictor	Correlation
1	Age	.029
2	Years of Education	.185
3	Defense Language Aptitude Test (DLAT)	.373
	Modern Language Aptitude Test (MLAT):	
4	MLAT - Part 1: Number Learning	.155
5	MLAT - Part 2: Phonetic Script	.307
6	MLAT - Part 3: Spelling Clues	.324
7	MLAT - Part 4: Words in Sentences	.359
8	MLAT - Part 5: Paired Associates	.243
9	MLAT - Auditory (Total 4 & 5)	.266
10	MLAT - Paper and Pencil (Total 6, 7, & 8)	.413
	MLAT - Total	.401
	Pimsleur Language Aptitude Battery (PLAB):	
12	PLAB - Part 1: Past Grades (Biographical)	.258
13	PLAB - Part 2: Interest	.113
14	PLAB - Part 3: Vocabulary	.267
15	PLAB - Part 4: Language Analysis	.263
16	PLAB - Part 5: Sound Discrimination	.198
17	PLAB - Part 6: Sound Symbol Assoc.	.204
18	PLAB - Auditory (Total 16 & 17)	.253
19	PLAB - Paper and Pencil (Total 14 & 15)	.319
20	PLAB - Linguistic (Total 18 & 19)	.357
21	PLAB - Total	.405
22	Otis - Lennon (IQ)	.272
23	Need for Social Approval Scale	-.078
24	Taylor Manifest Anxiety Scale	-.040
25	Defense Language Aptitude Battery	.431

MONTE CARLO COMPUTER PROGRAMS FOR SIMULATING SELECTION DECISIONS FROM PERSONNEL TESTS

MR. JOHN W. THAIN
TESTS AND MEASUREMENTS
EVALUATION SPECIALIST

A PAPER PRESENTED TO THE
MILITARY TESTING ASSOCIATION
OCTOBER 1978

DEFENSE LANGUAGE INSTITUTE
FOREIGN LANGUAGE CENTER
PRESIDIO OF MONTEREY, CA 93940

ABSTRACT

From a strictly theoretical point of view, the best method for determining the cutoff score for an aptitude test is to administer the test during a field trial, but randomly select personnel for training regardless of score on the aptitude test, and then observe performance and attrition rates in terms of aptitude test score. However, such a method is often impractical in a military setting. Monte Carlo techniques can be used to simulate random selection from the "unrestricted" population to which the aptitude test is administered. A computer program designed by the author uses test and criterion parameters and cutting scores, correlation coefficient, sample size, and number of samples to be drawn as inputs, and calculates decision classification rates across samples and for combined samples.

MONTE CARLO COMPUTER PROGRAMS FOR SIMULATING SELECTION DECISIONS FROM PERSONNEL TESTS

My colleague at DLIFLC in Monterey, Mr. Henderson, has presented a paper on the development of the Defense Language Aptitude Battery (DLAB). My work on a Monte Carlo program for simulating personnel test decisions was done in conjunction with the development of this language aptitude test. The Monte Carlo program has broader application for all personnel tests used for selection, not just our language aptitude test. However, our work on DLAB provides concrete illustrations of how the program can be used. I will start by explaining some relatively basic concepts and terms and build up to more complicated ideas.

Since I am not sure about the background of the audience, I am not sure how to sequence the presentation of these basic concepts. I hope the main ideas eventually get through and that no one feels distressed because he has difficulty in following the transitions from point to point.

The table at Appendix A illustrates a basic problem in making personnel decisions. No language aptitude test is perfect enough to insure that everyone scoring higher than a certain score will succeed in language training while everyone scoring lower will fail. In practice, some examinees pass the test and then fail the training. Other examinees fail the test, but could have passed the training if they had been given the chance.

If the minimum passing score is extremely low, then we will have a preponderance of problems of the first kind -- examinees passing the test but subsequently failing training, but very few problems with screening out nonpassing examinees who could have succeeded in training. If the cutting score is very high, we will have only minor problems with test passers subsequently failing training but more significant problems with the unwanted screening out of nonpassers who failed to achieve the high passing score but who could have succeeded in training.

Let us call the first type of error (where test passers fail in training) false positive errors, and the second type of error (where examinees are screened out who could have succeeded in training) false negative errors. As we can see, the proportion of false positives will increase as the cutoff score falls, and the proportion of false negatives will increase as the cutting score rises. By choosing a given cutting score we choose a certain tradeoff between false positives and false negatives.

Let us assume a simple scenario involving a predictor test and a criterion test. All examinees taking the predictor test also take the criterion test. We want to generalize these results to a future situation in which we have established cutoff scores for the predictor test and a cutoff score for mastery on the criterion test.

Let us make some assumptions that will allow us to conduct a simulation study. Assume that we know the population parameters listed on the slide -- mean and standard deviation of predictor and criterion tests and the correlation between predictor and criterion. Assume a bivariate normal distribution. Our computer program can easily utilize a random number generator and then an inverse normal distribution function to generate a normal distribution of mean μ_X and standard deviation of σ_X. A corresponding set of random numbers can be generated and subjected to the inverse normal distribution function and then plugged into the formula in Appendix B.

This formula will generate a criterion distribution with a mean of μ_y, standard deviation of σ_y and correlation of r_{xy} with the predictor. If z scores are used throughout the computation, the formula is much simpler. It was also simpler to write the computer program using z scores and convert to raw scores only when output was required.

So we now have a bivariate normal distribution with given parameters. At least we almost have such a distribution. No system for generating random numbers is totally random. The random number functions I have used have a very slight bias. The more pairs of random numbers generated, up to a certain point, the closer the distribution generated approaches population parameters. Since we are dealing with a computer simulation, we can easily generate 50,000 or more pairs of random numbers, and the amount of bias is extremely small.

Appendix C is an output from a computer program I have written. Note the parameter values of 43.6, 69.3, 18.0, 14.6, and .63, and the actually obtained values of 43.4969, 69.27282, 17.98392, 14.56758, and .63320. Corresponding z values are very close to a mean of 0 and standard deviation of 1 as we would expect: -.00573, -.00186, .99911, and .99778. By using the random number function and the inverse normal distribution function, we come very close to a bivariate normal distribution with the desired parameters.

The next step in the simulation process is to add the considerations of predictor cutting score and criterion cutting score we mentioned earlier. Every predictor score is matched to a criterion score. We mentioned earlier two undesirable combinations of predictor and criterion scores, which we labeled false positives and false negatives. Two other desirable combinations exist -- combinations in which the aptitude test is doing what it is supposed to do. One combination is when the examinee passes the test and then passes his training; we call these cases valid positives. Another combination is that the examinee fails the test and would have failed the training also; we call these cases valid negatives.

Given our bivariate distribution and our predictor and criterion cutoffs, every pair of predictor and criterion scores falls into one of these four categories. As our computer program generates its bivariate distribution, it simultaneously counts the number of cases in each of the four categories, given the cutoffs specified.

The results are shown in the computer printout at Appendix D. In this case our predictor cutoff of 60 is almost a full standard deviation above the predictor mean of 43.6. The computer program generated 25,000 cases, 4502 of which passed the predictor test and 20,498 of which failed the predictor test. As explained earlier, the "passes" are further divided into valid and false positives and the "fails" into valid and false negatives, depending on whether the criterion cutoff score was achieved. In this printout, for a given predictor and cutoff score we have the number of positive and negative cases and the number of cases in each quadrant. We also have the mean predictor and criterion score in each of these categories.

If the distribution stays constant as the predictor cutoff rises, fewer people pass. Some of the correctly classified valid positives that were below the cutoff become misclassified as false negatives. However, some of the false positives that were below the cutoff become properly classified as valid negatives. This is the kind of tradeoff we mentioned earlier.

It is interesting to note that as the predictor cutoff rises in this example, the mean scores for both predictor and criterion in all quadrants rise. This occurs because there is a relatively high correlation between predictor and criterion. Increasing the predictor cutoff adds cases with a higher predictor score to the category of negative cases, so that both the predictor and criterion means for negative cases rise. The same cases that are added to the negative categories were the lowest predictor scores from the positive categories so that the predictor and criterion means for the positive categories also rise for the remaining cases with higher predictor scores. In parentheses on the printout we also have the percentage of positive and negative cases passing the criterion for each of the predictor cutoffs.

We have generated 25,000 cases so that the bivariate distribution would have almost exactly the parameters desired. However, at DLIFLC we don't admit students to training in groups of 25,000. Our computer program enables us to break the big sample into a number of smaller samples of any size we choose. This enables us to view the effect of sampling error in circumstances similar to everyday operating conditions. In our case we have drawn 250 samples of 100 each from the 25,000 cases. To establish a frame of reference for thinking about sampling error, we have printed the standard errors of the parameters for 25,000 cases and for 100 cases in the first two columns at the top and the left of the output page in Appendix C.

At the lower left of the page we have the mean values of the small sample characteristics across groups. Of course the mean of the means is the same as the grand mean, but the mean of the other characteristics varies slightly from the values for the whole sample. On the lower right we have the standard deviation of these characteristics across groups. For example, with this predictor grand mean and this standard deviation of predictor sample means, we would expect to find 68% of the sample means to fall between about 41.7 and 45.3.

Finally, for each combination of predictor and criterion cutoff scores there is an output page like Appendix E. This page gives the average number of cases in each of the small samples for each of the four categories and also the standard deviation of the number of cases across all the small samples. For purposes of comparison, a page is shown where the predictor cutoff has been raised from 60 to 68.

In the preceding example, we have only talked about changing the predictor cutoff. It is just as feasible to change the criterion cutoff, and such examples were not shown in order to keep the presentation simple.

In summary, this program employs a Monte Carlo technique to generate a bivariate normal density function. The five parameters on which the function is based are the predictor and criterion means, the standard deviations, and the correlation coefficient. The program treats these parameters as population values from which repeated samples are drawn. Individual cases are then compared to predictor and criterion cutting scores; the rates and distributions of valid positives, false positives, valid negatives, and false negatives are then computed.

The predictor and criterion cutting scores can be automatically incremented to produce expectancy tables. The program utilizes a random number generator as input to an inverse normal function taken from STATPAK (Computer Sciences Corporation, 1972) to create the test and criterion distributions. It has been adapted for a UNIVAC 1108 with a Fortran V Compiler.

The following output is generated:

1. Standard errors of parameters.

2. Obtained means, standard deviations, and correlation coefficient for combined sample.

3. Standard deviation of means, standard deviations, and correlation coefficients across samples.

4. As relates to the four decision categories (VP, FP, VN, FN):

a. Average number of cases in combined samples.

b. Average number of cases in each sample.

c. Average percentage of cases.

d. Standard deviation of number of cases across samples.

e. Predictor mean score.

f. Criterion mean score.

5. Proportion of successful selectees.

6. Mean predictor score for selectees.

7. Mean criterion score for selectees.

8. Proportion of successful rejections (assuming rejections were given the opportunity to succeed).

9. Mean predictor score for rejections.

10. Mean criterion score for rejections.

APPENDICES

APPENDIX A

COMPUTER SIMULATION OF DLI SELECTION PROBLEM

DLAB CUTTING SCORE	BASE SAMPLE INELIGIBLE %	ATTRITION %	AVERAGE GRADE %
40	42.2	17.8	78.7
42	46.5	16.7	79.1
44	50.7	15.3	79.6
46	54.6	14.0	80.0
48	58.8	12.7	80.5
50	62.9	11.5	81.0
52	66.9	10.5	81.4
54	70.6	10.0	81.9
56	74.0	8.6	82.3
58	77.2	7.7	82.8
60	80.3	7.1	83.3
61	81.7	6.7	83.6
62	83.1	6.1	83.9
63	84.4	5.6	84.2
64	85.5	5.2	84.4
65	86.7	4.8	84.7
66	87.9	4.4	85.0
67	88.9	4.0	85.2
68	90.0	3.9	85.5
69	90.1	3.5	85.9

APPENDIX B

$$y_i = \mu_y + \sigma_y r_{xy} \left(\frac{x_i - \mu_x}{\sigma_x} \right) + K_i \sigma_y \sqrt{1 - r_{xy}^2}$$

μ_y = CRITERION MEAN

x_i = PREDICTOR SCORE

μ_x = PREDICTOR MEAN

σ_x = PREDICTOR STANDARD DEVIATION

σ_y = CRITERION STANDARD DEVIATION

r_{xy} = CORRELATION BETWEEN PREDICTOR AND CRITERION

K_i = A NUMBER GENERATED BY UTILIZING A RANDOM NUMBER GENERATOR AND INVERSE NORMAL DISTRIBUTION FUNCTION TO GENERATE A DISTRIBUTION WITH $\mu_K = 0$ AND $\sigma_K = 1$.

$\sigma_y \sqrt{1 - r_{xy}^2}$ = STANDARD ERROR OF ESTIMATE

WITH Z SCORES

$$y_i = r_{xy} z_x + K_i z_{\sigma_{yx}}$$

σ_{yx} = STANDARD ERROR OF ESTIMATE

APPENDIX C

```
                         MONTE CARLO SIMULATION OF PERSONNEL DECISIONS

              STANDARD ERRORS OF PARAMETERS

                                                                             LARGE SAMPLE           SMALL SAMPLE
PARAMETER         STANDARD ERROR     STANDARD ERROR   PARAMETER           RAW           Z              RAW
                  LARGE SAMPLES      SMALL SAMPLES      VALUE

PREDICTOR MEAN        .11384            1.80000       43.60000         43.49690     -.00573         43.49690
CRITERION MEAN        .09234            1.46000       69.30000         69.27282     -.00136         69.27282
PREDICTOR S.D.        .08050            1.27279       18.00000         17.98392      .99911         17.94722
CRITERION S.D.        .06529            1.03238       14.60000         14.56758      .99718         14.52660
CORRELATION R         ******            ******          .63000           .63320     ******           .63066
            Z         .00632            .10153          .74142           ******      .74674           .74965

                                     TOTAL SAMPLE    25000.
                                     SMALL SAMPLE      100.

                                                      STANDARD DEVIATION OF SMALL SAMPLE CHARACTERISTICS
   SMALL SAMPLE MEAN CHARACTERISTICS
                                                      PREDICTOR MEAN ACROSS GROUPS    1.79909
PREDICTOR MEAN ACROSS GROUPS       43.49690           CRITERION MEAN ACROSS GROUPS    1.45945
CRITERION MEAN ACROSS GROUPS       69.27282           PREDICTOR S.D. ACROSS GROUPS    1.24769
PREDICTOR S.D. ACROSS GROUPS       17.94722           CRITERION S.D. ACROSS GROUPS    1.03659
CRITERION S.D. ACROSS GROUPS       14.52660           CORRELATION ACROSS GROUPS        .06356
CORRELATION MEAN ACROSS GROUPS       .63066           FISHER'S Z ACROSS GROUPS         .10707
FISHER'S Z ACROSS GROUPS             .74965
```

```
1                                    DATA ON INDIVIDUAL QUADRANTS                                                      1
      CUTOFFS,              CATEGORY   N           OVERALL MEANS               MOMENTS ACROSS SMALL SAMPLES
 PREDICTOR  CRITERION                            PREDICTOR    CRITERION           SKEWNESS          KURTOSIS
  59.00000   70.00000          VP    4069.      69.21481     85.83513             .402390           .183520
                               FP     751.      65.75937     63.77921             .561168           .091403
                               FN    7892.      44.28221     79.16371             .209500          -.028825
                               VN   12208.      32.54277     57.77206            -.143524           .456057
                                P    4820.      69.10673     82.33270( .844)
                                N   20180.      37.38003     66.13779( .391)


  60.00000   70.00000          VP    3838.      70.43920     85.97907             .451806           .402362
                               FP     664.      66.01269     63.81046             .662683           .462120
                               FN    8123.      44.72053     79.28511             .245859          -.082857
                               VN   12375.      33.12955     57.81266            -.144871           .485821
                                P    4502.      69.78685     82.70942( .853)
                                N   20498.      37.72285     66.32182( .396)


  61.00000   70.00000          VP    3583.      71.14897     86.16845             .327903           .380445
                               FP     564.      66.99305     63.86471             .794985           .635429
                               FN    8378.      45.20007     79.40785             .241509          -.094557
                               VN   12475.      33.34882     57.85830            -.149963           .391269
                                P    4147.      70.58376     83.13510( .864)
                                N   20853.      38.11023     66.51615( .402)


  62.00000   70.00000          VP    3333.      71.87426     86.43228             .337208          -.024424
                               FP     468.      67.85481     64.02303             .749176           .516461
                               FN    8628.      45.67178     79.50182             .180784          -.080955
                               VN   12551.      33.51903     57.88451            -.163554           .320410
                                P    3821.      71.30091     83.57027( .872)
                                N   21179.      38.46987     66.69344( .407)


  63.00000   70.00000          VP    3094.      72.59835     86.69208             .350595           .010669
                               FP     425.      68.64796     64.11780             .908911           .741669
                               FN    8867.      46.12538     79.59796             .120242          -.150939
                               VN   12614.      33.66330     57.91596            -.235581           .342229
                                P    3519.      72.12125     83.96572( .879)
                                N   21481.      38.80773     66.86593( .413)
```

```
                    PREDICTOR CUT-OFF 60.00000
                    CRITERION CUT-OFF 70.00000

                         VALID POSITIVES

NUMBER OF CASES IN TOTAL SAMPLE                                           3838.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                          15.35200
AVERAGE PERCENTAGE OF CASES                                             .153520
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES                3.57761
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES          .035776

                         FALSE POSITIVES

NUMBER OF CASES IN TOTAL SAMPLE                                            664.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                           2.65600
AVERAGE PERCENTAGE OF CASES                                             .026560
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES                1.59871
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES          .015987

                         FALSE NEGATIVES

NUMBER OF CASES IN TOTAL SAMPLE                                           8123.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                          32.49200
AVERAGE PERCENTAGE OF CASES                                             .324920
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES                4.34574
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES          .043457

                         VALID NEGATIVES

NUMBER OF CASES IN TOTAL SAMPLE                                          12375.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                          49.50000
AVERAGE PERCENTAGE OF CASES                                             .495000
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES                4.86711
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES          .048671
```

```
                     PREDICTOR CUT-OFF 68.00000
                     CRITERION CUT-OFF 70.00000

                          VALID POSITIVES

NUMBER OF CASES IN TOTAL SAMPLE                                      1987.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                      7.94800
AVERAGE PERCENTAGE OF CASES                                        .079480
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES           2.62083
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES     .026208

                          FALSE POSITIVES

NUMBER OF CASES IN TOTAL SAMPLE                                       184.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                       .73600
AVERAGE PERCENTAGE OF CASES                                        .007360
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES            .87025
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES     .008702

                          FALSE NEGATIVES

NUMBER OF CASES IN TOTAL SAMPLE                                      9974.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                     39.89600
AVERAGE PERCENTAGE OF CASES                                        .398960
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES           4.62892
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES     .046289

                          VALID NEGATIVES

NUMBER OF CASES IN TOTAL SAMPLE                                     12855.
AVERAGE NUMBER OF CASES IN EACH SMALL SAMPLE                     51.42000
AVERAGE PERCENTAGE OF CASES                                        .514200
STANDARD DEVIATION OF NO. OF CASES ACROSS SMALL SAMPLES           4.79830
STANDARD DEVIATION OF PERCENTAGE OF CASES ACROSS SMALL SAMPLES     .047983
```

SECTION 8

METHODS OF DETERMINING PERSONNEL AVAILABILITY

UNCLASSIFIED

PAM: A METHODOLOGY FOR PREDICTING AIR FORCE PERSONNEL AVAILABILITY

BY

H. Anthony Baran
Duncan L. Dieterly

Advanced Systems Division
Air Force Human Resources Laboratory
Wright-Patterson Air Force Base, Ohio 45433

Andrew J. Czuchry
John C. Goclowski
Fredric F. Phillips
Stuart E. Peskoe
Anthony J. LoFaso

Dynamics Research Corporation
Wilmington, Massachusetts

651

PAM: A METHODOLOGY FOR PREDICTING AIR FORCE PERSONNEL AVAILABILITY

ABSTRACT

This paper describes a methodology for projecting the career transition activity of Air Force personnel to predict their future availability. It includes a personnel availability analysis model (PAM), application techniques, and a personnel data bank.

The cost significance of weapon system personnel requirements has made their consideration a major concern within the systems acquisition process. The need for tools to aid in this consideration has led primarily to the development of models and techniques which address the identification of those requirements. What has been lacking is a means to determine and provide guidance for the accommodation of the potential impacts of a changing military force structure on their fulfillment. The methodology described here is a first step toward the comprehensive assessment of weapon system design, personnel requirements, and support plans in terms of the future availability of military personnel.

The heart of the PAM methodology is a computerized model which represents career transition activity within the Air Force by a series of Markov processes, each depicting a subpopulation of airmen, with states defined by years of service and paygrade. State transition probabilities are calculated on the basis of actual transition activity data contained in the Uniform Airman Record (UAR). Subpopulations may either be defined on an a-priori basis, such as by Air Force Specialty Code (AFSC) designation, or analytically established by applying a discrete dependent variable regression analysis technique called Logit Analysis. This technique identifies subpopulations consisting of

personnel exhibiting similar career transition behavior and describes them in terms of individual attribute data contained in the UAR. It increases career projection accuracy by reducing uncontrolled variance, and provides increased specificity in the analysis of personnel policy change impacts.

The PAM methodology includes computer programs which extract and combine data elements from the UAR to form an addressable data bank. Presently, this data bank contains a selection of data elements from the 1975, 1976, and 1977 UAR files for approximately 95,000 airmen assigned to thirteen AFSCs.

BACKGROUND

The cost and quality of trained system support personnel have become extremely important considerations in weapon system design and support planning. The primary reason for attributing such importance to the role of human resources in weapon system development is the growing concern that their cost, which presently overshadows that of system acquisition, will grow to a size which will effectively preclude the affordability of future systems. The Air Force, in particular, as a branch of the Armed Services whose operational effectiveness is most often measured in terms of the capabilities of its weapon systems, has the unfortunate distinction of being the Service most likely to experience the object of that concern.

This situation has precipitated considerable research concerning the development of tools and techniques to implement the consideration of human resources implications of design, operation and support within the systems acquisition process. Emphasis has been placed, however, on human resources as requirements rather than as vital commodities whose availability and operational disposition over time may be as crucial to new weapon system deployment as is their timely specification as system ownership requirements. (Slide 1)

The personnel availability methodology which this paper describes is one product of an effort which expands that emphasis to address: (1) the present and future capability of a military personnel force structure to respond to those requirements; and (2) how that structure may be perturbed to increase its ability to do so. That effort is the Air Force Human Resources Laboratory Project 1959, entitled "Advanced System for Human Resources Support of Weapon System Development." It was undertaken to develop a coordinated human resources technology package which combines the results of previous research in several human resources related technologies to provide a single integrated mechanism for the use of human resources considerations as system design and support planning guidelines. The overall objective is to avoid unnecessary system ownership cost through a methodical consideration of all aspects of personnel, manpower, and training within the design process itself.

Much of the integrated technology package addresses the early assessment of system design and support alternatives in terms of their potential impact on human resources requirements. However, that portion which constitutes the personnel availability model (PAM) methodology complements that activity by providing further guidance concerning the feasibility of meeting those requirements within the constraints imposed by present and foreseeable circumstances of personnel availability. It allows system planners the option of either designing a system in compliance with a predicted personnel availability situation or seeking means to alter that situation to provide for a mission essential design capability. In the former instance, the PAM methodology can probabilistically define the composition of the personnel force structure at the time of system deployment, thus identifying a design/planning requirement. In the latter, it can provide a vehicle for rapid hypothesis testing in a search for that set of personnel policy actions most likely to result in the availability of appropriate personnel to meet mission essential support requirements.

The modeling approach to personnel availability analysis embodied in the PAM methodology is neither new in concept nor uniquely superior to others in terms of its ability to provide panacea-like solutions to insurmountable problems. However, as a total package of model, data base, and application techniques, it represents an important first step toward the comprehensive assessment of weapon system design, personnel requirements, and support plans in terms of the means available to operationally accommodate them.

OPERATION OF THE MODEL

The heart of the PAM methodology is the personnel availability model itself. Its objective function is to provide estimates of future personnel availability on the basis of career transition activity projections derived from historical data indicating current and past force structure composition and past career transition activity. Derived primarily to meet needs identified for guidance within the Air Force systems acquisition process, that function is predicated on the following five capability requirements: (1) evaluation of the current human resources in the Air Force; (2) estimation of that human resources complement at future points in time; (3) comparison of estimated human resources availability to estimated requirements at coincident points in time; (4) quantification of differences between human resources requirements and estimated availability; and (5) identification of personnel policy changes necessary to reduce or eliminate potential disparities between future personnel requirements and future personnel availability.

It was originally expected that a model could be selected or adapted from among the many manpower/personnel models which exist today. To a certain extent that expectation was borne out. However, additional operating requirements were identified which extended the modeling capability requirements of this effort beyond those of existent candidate models identified in an extensive literature search. These requirements called for: the identification and tracing of actual personnel career transition activity; the consideration of management conditions within the manpower system, such as training and retirement policies; the calculation and use of probabilistic information; minimization of computational requirements without substantial loss in capability to accurately reflect the actual functioning of the Air Force manpower system; and operational specificity sufficient to assess the career transition activity of subpopulations within the total Air Force personnel population, defined by personnel attribute designations, while maintaining enough flexibility to investigate larger aggregate populations. In addition, the model to be selected had to be capable of projecting the future size of the personnel complement to be found within a subpopulation category, itself defined either by personnel attribute or career status designations.

In order to meet the previously defined modeling requirements and to provide a realistic representation of the Air Force manpower system, the career transition process of Air Force personnel is most tractably modeled as a finite-state, discrete-time Markov process. This conclusion was reached in consideration of the following aspects of the Air Force manpower system which are compatible with such a model: (Slide 2)

1) It is hierarchical when states are defined by years of service (YOS) and paygrade. Airmen can only move from low paygrades and low years of service to higher paygrades and higher years of service, if they are to remain in the system.

2) It is approximately Markovian. An airman's state (YOS, paygrade) at time t+1 depends primarily on his state at time t, and less so on his state at prior times.

3) It is discrete. An airman transitions from one state to another at yearly intervals, rather than at random time intervals.

A Markov model is structurally suited to take advantage of the above three properties of the Air Force manpower system, and also has the virtue of being computationally facile. These facts are underscored by the mechanical simplicity of the Markov model chosen to meet the requirements of this effort.

A population possessing user specified attributes is partitioned by the model into a state matrix, with states defined by

YOS and paygrade. (Slide 3) Once the state matrix for a given population has been defined, the model computes the probabilities associated with the various allowable types of state transition. This is accomplished on the basis of two sets of historical data abstracted from the Uniform Airman Record (UAR). The data is sampled at two points in time, the interval between which is determined by the projection interval which a user desires to be produced by the model. In the present case, a one year model projection interval was desired. Therefore, the UAR was sampled at two points in time one year apart (1975 and 1976). It should be noted here that the model' s overall prediction of personnel availability at a future point in time is accomplished on the basis of an iterative updating of its state population projection. That is, the model continually applies the UAR data-derived transition probabilities (calculated from the actual transitions indicated by the two point data sample) to its most recent state population projection matrix, until the user specified outyear termination point is reached. The final state population projection matrix is the airman population prediction for that outyear bounded, of course, by whatever personnel attribute or career status designations the user has chosen to impose as output constraints.

Several assumptions were made concerning both the flow of airmen through the Air Force manpower system and external policy considerations which might conceivably affect the probabilities associated with various types of career transition activity. (Slide 4) In formulating the model, it was assumed that once a person enters a particular state the probabilities associated with his next transition are independent of how or from where he may have arrived at that state. That is, the probability associated with his next transition must be selected from those available to any other person in that same state. The second model formulation assumption is that a transition must occur within the model projection interval. The third is that transitions must indicate progress either in years of service or paygrade, i.e., no demotions are allowed and time in service must increase without regard for the actual but rare incidence of breaks in tenure. The fourth assumption postulates a constant recruitment rate. This is to keep the transition probabilities "clean" with respect to variables other than those under examination. As will be explained later, any of these assumptions may be purposively violated by the user by exercising the user-interactive features of the PAM.

A typical state within the model is illustrated in Slide 5. Such a state represents a particular service tenure and paygrade within a single Air Force Specialty Code (AFSC), and may be visualized as a single cell within a three dimensional state population projection matrix bounded by those variables. It is so represented within the PAM, with the additional consideration that the population of each state is also bounded by the set of personnel attribute constraints imposed by the model user. In the illustration, the solid arrows indicate the transition probabilities that produce a new state population (S_{ij}'). Segmented arrows indicate the ways in which personnel may leave a typical state. As is shown, transition into a state can occur only in one of three ways: (1) a transfer from some other AFSC, or a new accession; (2) an increase in paygrade with an incremental increase in YOS; or (3) an incremental increase in YOS without a change in paygrade.

Basically, the PAM examines its historical data base and calculates the exit probabilities associated with each state as determined by historical precedent. It then forms a state and probability data base which is comprised of transition probability matrices for upgrade, increment, loss from service, and transfer. These matrices become the bases for determining the composition of future state populations. The basic Markov formulation for these future state probability calculations is shown in Slide 6. The following provides additional detail.

Letting S_{ij} denote a state at time t with year of service i and paygrade j, and S'_{ij} denote the state population at time $t+1$, the following equation defines the computation for determining a new state value for the succeeding time interval:

$$S'_{ij} = (S_{i-1,j})(PI_{i-1,j}) + (S_{i-1,j-1})(PU_{i-1,j-1}) + T_{ij}$$

where:

$S'ij$ = the state population at a point in time $t+1$, having i years of service and j paygrade;

$Si-1,j$ = the state population at a point in time t, having $i-1$ years of service and j paygrade;

$Si-1,j-1$ = the state population at a point in time t, having $i-1$ years of service and $j-1$ paygrade;

$PIi-1,j$ = the probability that people in state $Si-1,j$ will increment one year in service but, will not leave the population or upgrade;

$PUi-1,j-1$ = the probability that people in state $Si-1,j-1$ will upgrade to $S'ij$ within the next time interval;

Tij = the number of people from outside the population that will transfer into state $S'ij$ within the next time interval.

It should be noted that, since year of service is monotonic, the i subscript carries time in the equation for the succeeding state. New state population values are determined by the transitions from the state preceding it by one time interval. Probability of loss from a state, by exit from the Service or transfer to another population, is not included in the previously described equation but is taken into consideration in the more comprehensive state transition equations within the PAM.

PAM DATA BASE

In the most general terms, the PAM data base can be described as being comprised of two sets of data which provide "snapshot" pictures of the Air Force personnel population at two points in time. Their comparison, within the model, reveals the career transition activity which has taken place within the time interval selected and provides the means to project that which will take place during successive time intervals in the future. The equation described above defines the process for that projection. The data bases are abstracts of the Uniform Airman Record (UAR). The current PAM data base was constructed from the 1975 and 1976 UAR and covers approximately 95,000 airmen in thirteen technical Air Force Specialty Categories. (Slide 7) Data for individual personnel are assembled on PAM records and address such things as test scores, duty assignments, personnel history, pay levels, etc. which the PAM uses as personnel descriptors or attributes. The records cover 24 of the 450 identifiable personnel characteristics contained in the UAR. (Slide 8) The present selection was made on a judgemental basis and could conceivably be improved for specific PAM application objectives by a more detailed consideration of the possible relationships between individual attributes and those objectives.

The actual process of UAR data abstraction, necessitated by the voluminous nature of the UAR, is undertaken external to the PAM. It should be considered a preparation process, rather than a PAM function, in that the proper selection of attributes to be abstracted demands considerable user judgement. In any case, the mechanization of the process, once selection decisions are made, is a straightforward task. In the present instance, it was rapidly accomplished by the Computation Sciences Division of the Air Force Human Resources Laboratory. Once the abstraction process is performed and the two PAM data record sets are compiled, the PAM takes them as inputs and combines them to form a single record set. This combined record set is then used by the PAM to generate transition probability data, i.e., future behavioral probability data based on the recorded career transition activity concerning whether individuals in certain AFSCs incremented in years of service, upgraded in pay status, did both, transferred to another AFSC, or left the Service within the time interval covered.

Transition probabilities are computed using the transition data and the following algorithms:

(1) Computation of state (S) matrix:

Sij = number of airmen in the population with years of service (i) and paygrade (j);

(2) Computation of probability (P) matrices

$$PUij = \frac{\text{Number of airmen who left state (i, j) during the time interval via upgrade}}{Sij = \text{population of state (i, j) at beginning of time interval}}$$

$$PIij = \frac{\text{Number of airmen whose grade remained the same during the time interval}}{Sij = \text{population of state (i, j) at beginning of time interval}}$$

$$PLij = \frac{\text{Number of airmen who left state (i, j) during the time interval by leaving the Service}}{Sij = \text{population of state (i, j) at beginning of time interval}}$$

Probabilities associated with airmen transfer in and out of Air Force Specialty Categories (AFSCs) are included among the probability matrices calculations within the PAM. The process by means of which they are generated involves a comparison of the four types of AFSC designations found in the UAR, viewed at the two points in time which bound the interval of historical data sampling. Discussion of that process is beyond the scope of this paper.

PAM PROGRAMS AND FUNCTIONS

The PAM is comprised of two computer programs which perform the following four functions: (1) data base generation; (2) data base maintenance; (3) extrapolation over time; and (4) data post-processing. (Slide 9)

Each function is performed by a program or program subroutine and the functions are sequential in the sense that the output of one function serves as the input to the next function in the sequence.

Program 1 (Data Base Generation Function)

This program selects the personnel records on the basis of user-selected criteria for screening on particular attributes. The records selected are processed and used to create matrices for use by subroutines two and three of program 2. Once the records have all been processed, the program calculates probabilities (P' s) of upgrade, increment in years of service, and loss for each AFSC, paygrade, and year of service state as was previously described. The program also accumulates matrices for numbers of transfers by AFSC, years of service, and paygrade.

Program 2 (Model Operation Function)

This program consists of three subroutines and is user-interactive via remote terminal facilities. It performs the following tasks: (1) data base maintenance which allows for user-entered override modifications to the state, transfer, and probability matrices; (2) operation of the probability extrapolator model for a user-specified number of time intervals; and (3) printout of results and/or current values contained in the state, transfer, or probability matrices. All of the above functions are controlled by the user at the computer terminal. Through responses to a series of questions displayed at the terminal, program execution will select subroutine 1, 2, or 3 of this program, depending upon the specific requirements of a task defined by the user. (Descriptions of these subroutines are given below.) Each time a user input is required, a statement followed by a question mark will be displayed to prompt the user. At those points, program execution will pause for the user response and resume once it is made. Termination of each user input is signaled to the program by striking the carriage return key. The three subroutines function as follows.

Program 2; Subroutine 1 (Data Base Maintenance Function)

This subroutine is used to modify the state, probability, or transfer matrices on an element-by-element basis. The subroutine will perform edit and reasonableness checks, i.e., allow user override of the matrix cell entries. Its use is optional and can be bypassed if the user is satisfied with the matrices computed from the input data supplied by program 1.

Program 2; Subroutine 2 (Extrapolation Function)

This subroutine produces the projections of state transfer, using the matrices calculated in Program 1. It steps the state matrix forward in time, interval by interval, and stores the output in a Result File for future output to the user in several ways and formats which he may designate. Examples of such outputs are: projections bounded by specified years of service, paygrades, and/or by specific outyear restrictions; and display via terminal screen or printed hard copy.

Program 2; Subroutine 3 (Post Processing Function)

This subroutine selectively lists any part of the Result File created by subroutine 2. The portion to be listed is a user option. Current programming allows the user to impose output parameter restrictions which yield listings within the following formats:

(1) The entire state matrix for paygrades 3 through 9 and years of service 1 through 21, where years of service 21 is an aggregation of years 21 through 30;

(2) Any selected combination of states; and

(3) Breakout by years of service with all paygrades being collapsed to yield a single line matrix output.

APPLICATION TECHNIQUES

Two very important features are built into the PAM. The first, provided by Program 2/Subroutine 1, allows the user to make changes in the state or probability matrices; thus giving him the opportunity to make state population projections on the basis of postulated, as well as real, initial state conditions and personnel career flow constraints. The second feature provides the PAM user with a capability to investigate the career transition activity of subpopulations selected on the basis of their members possessing certain combinations of personnel attributes, (Slide 10) e.g., age, education, race, sex, etc. Personnel availability investigation on the subpopulation level is desirable because the Air Force manpower system, while not truly a homogeneous system, has been modeled within the PAM as a Markov process. Specifically, various subpopulations of airmen have been found to have different career transition rates. Therefore, an increase in modeling accuracy may be obtained by dividing subject populations into homogeneous subpopulations for individual examination.

The PAM is structured such that projections are made on the basis of years of service (YOS) and paygrade. However, other factors may significantly affect the career transition process. Airmen with identical YOS and paygrade at present may be expected to transition to different states, independently of how they arrived at the current state, if there are sufficient differences among their other attributes such as test scores, marital status, or sex. Such differences are handled in the PAM methodology by two distinct mathematical approaches to the detailed evaluation of human resources availability in terms of their grouping on the basis of personnel attributes. Possible modeling inaccuracies which might arise as a result of attribute heterogeneities are accounted for by the calculation of separate transition rates for each homogeneous subpopulation that can be identified by attributes other than YOS and paygrade. Such descriptors may be thought of as driving attributes.

The first approach, which can be thought of as an attribute identification or categorical approach, employs a qualitative discriminant analysis using graphical displays of transition frequencies versus attributes. It identifies the attribute or combination thereof which indicates a high concentration of specific transitions. Directed at determining which groups of airmen, or subpopulations transition alike, it attempts to flag any attributes that are capable of being used to distinguish between airmen subpopulations exhibiting dissimilar career transition rates. This is accomplished by exposing incidents of career transition activity characterized by disproportionate numbers of transitions relative to the number of airmen possessing a given attribute or set of attributes. The resultant grouping of airmen reflects a categorization on that basis. The groups are then examined as individual entities which constitute homogeneously transitioning subpopulations.

The second approach is directed at the determination of a functional relationship between specific transitions and the attributes; as in multiple regression analysis. A specialized statistical technique called Logit Analysis is used to aid in that determination by establishing dependency relationships among the various attributes. The object of that analysis, rather than the identifcation of homogeneous groups of airmen in terms of career transition rates, is the determination of how the transition rates are related to the specific attributes themselves. In this formulation, transitions are viewed as a response (dependent) variable and the UAR data are taken to constitute a vector of explanatory (independent) variables. The variables are operated upon by a mathematical model, using binomial logit analysis, that in effect regresses the dependent variables on the independent variables to yield indicants of relationship. Results of this analysis are particularly well suited toward the provision of aid in the identification of airmen groups that have a distinct set of attributes and a set of transition probabilities different from the average movement parameters of the total airmen population, i.e., homogeneous subpopulations.

In summation, categorical analysis involves the identification of an attribute(s) that coincides with non-uniform population transition rates. It constitutes a search for driving attributes to define similarly transitioning subpopulations by comparing the frequency distributions of transitions. Logit analysis is directed towards the establishment of a weighting scheme for the attributes relative to transition rates. The process of subpopulation identification, state & probability matrices calculations, and human resources availability projection is functionally illustrated in Slide 11. Slide 12 provides a detailed illustration of PAM operation, to include the PAM methodology for identifying homogeneous subpopulations. The application methodology described above allows the subpopulation selection features of the basic PAM to be used intelligently on the basis of data implications, rather than that of trial and error. There is, however, a variable which, although not directly addressed by the PAM methodology, does bear significantly on the validity of the PAM personnel availability projections. That variable is recruitment rate. Normally, it may be assumed to be constant. However, the PAM data base maintenance function subroutine may be used to input changes at the discretion of the user.

Slide 13 illustrates an additional use for the identification of homogeneous subpopulations, other than prediction of total population availability on the basis of aggregating predictions for the subpopulations which it subsumes. That additional use is in the analysis of the potential impacts of personnel policy changes on career transition activity and future availability. It should be noted that, although the PAM provides a basic capability for this kind of analysis, the analysis of personnel policy impact assumes a knowledge of the relationships between policy and the personnel attributes which the PAM uses to track personnel career progression.

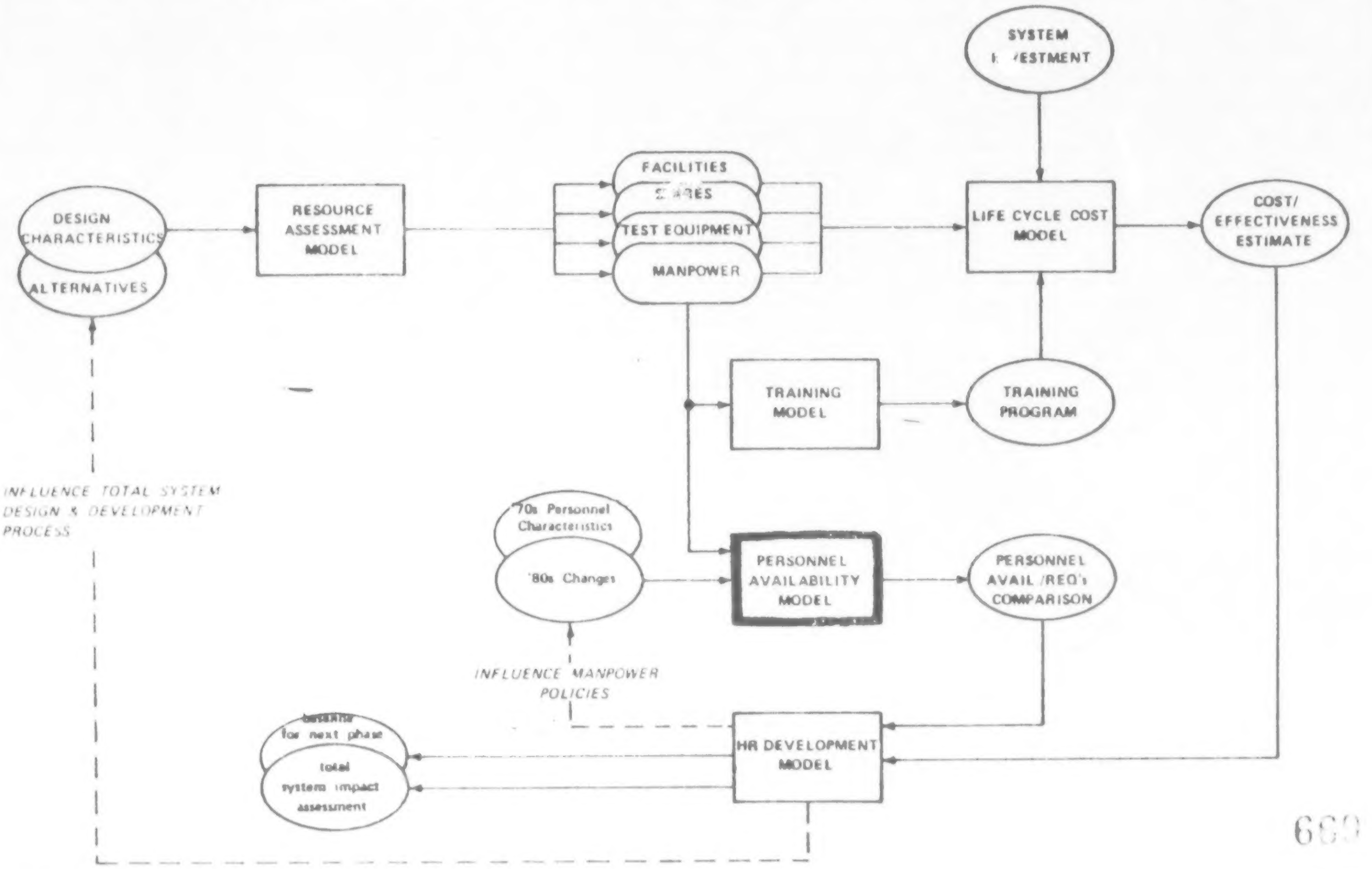
SUMMARY

A model has been developed for projecting the future availability of Air Force maintenance personnel based on the 1975 and 1976 career transition activity of 95,000 airmen. The utility and specificity of the model, and the accuracy of the projection results, have been expanded by the development of an application methodology. That methodology incorporates certain statistical techniques to, not only examine the Air Force manpower structure on a more individual basis than was heretofore possible but also to, identify personnel attributes that are related to tenure in the Air Force.

The PAM and its methodology for personnel availability analysis is capable of meeting the needs of an analyst who, in the design phases of weapon system acquisition, desires to determine whether planned maintenance manpower requirements may be expected to be fulfilled by the Air Force human resources supply at the time of weapon system implementation. If human resources availability projections indicate that requirements may not be met, the PAM provides a capability which can be of aid in seeking personnel policy changes which can be implemented to effect changes in the future availability of the required personnel such that future requirements can be met.

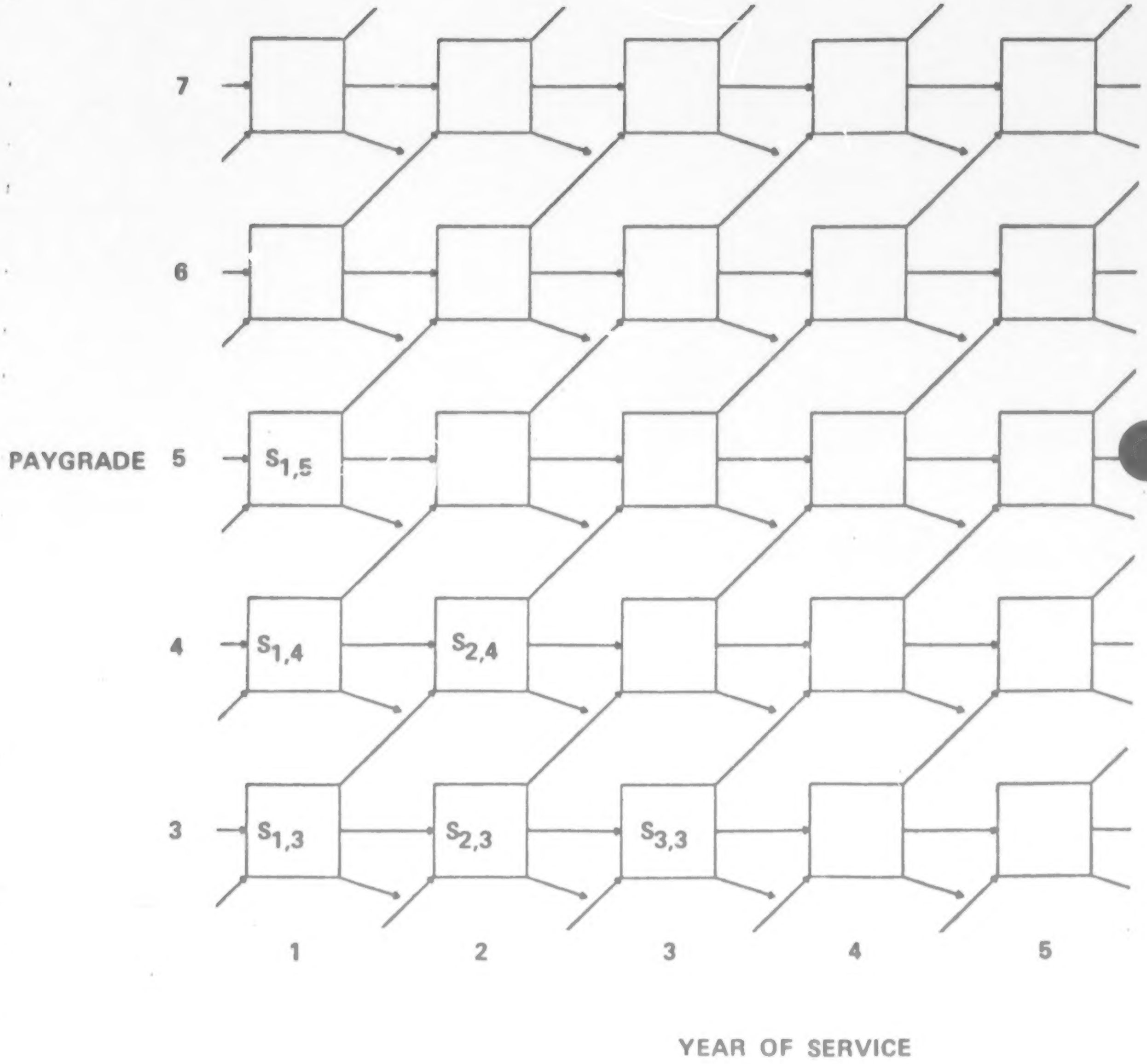
SLIDE 1

AFHRL ANALYTIC TOOLS


SYSTEM
INVESTMENT
DESIGN
CHARACTERISTICS
ALTERNATIVES
RESOURCE
ASSESSMENT
MODEL
FACILITIES
SPARES
TEST EQUIPMENT
MANPOWER
LIFE CYCLE COST
MODEL
COST/
EFFECTIVENESS
ESTIMATE
TRAINING
MODEL
TRAINING
PROGRAM
INFLUENCE TOTAL SYSTEM
DESIGN & DEVELOPMENT
PROCESS
'70s Personnel
Characteristics
'80s Changes
PERSONNEL
AVAILABILITY
MODEL
PERSONNEL
AVAIL./REQ's
COMPARISON
INFLUENCE MANPOWER
POLICIES
baseline
for next phase
total
system impact
assessment
HR DEVELOPMENT
MODEL

SLIDE 2

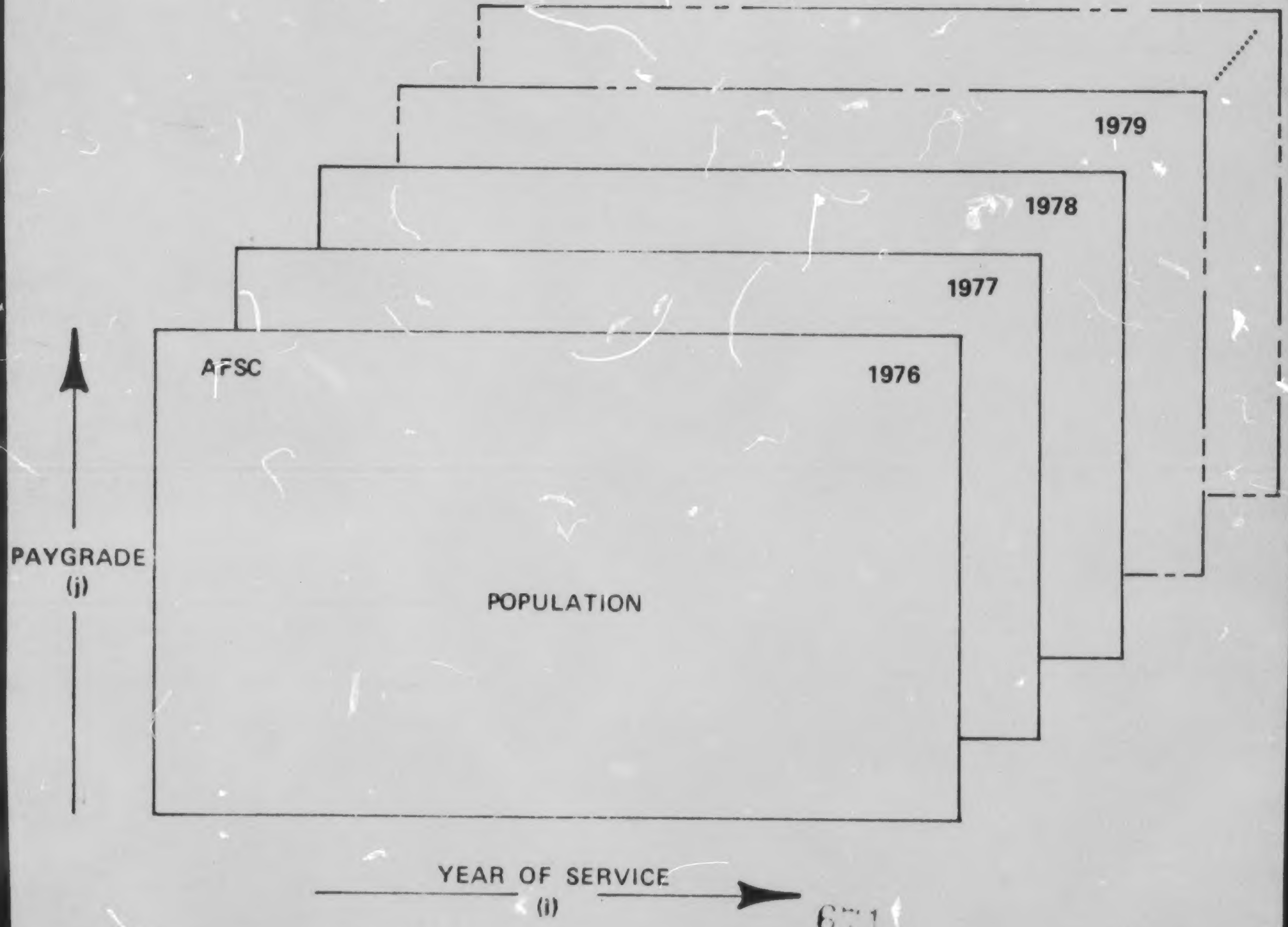
AIR FORCE
MANPOWER SYSTEM FLOW



7
6
PAYGRADE 5
4
3
$S_{1,5}$
$S_{1,4}$
$S_{2,4}$
$S_{1,3}$
$S_{2,3}$
$S_{3,3}$
1
2
3
4
5
YEAR OF SERVICE

SLIDE 3

POPULATION PROJECTION



AFSC

1976

1977

1978

1979

POPULATION

PAYGRADE
(j)

YEAR OF SERVICE
(i)

SLIDE 4

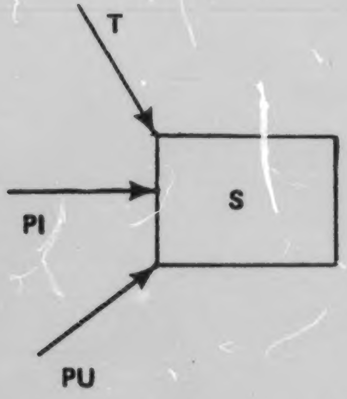
ASSUMPTIONS

FORMULATION

- PATH INDEPENDENCE
- HIERARCHICAL STRUCTURE

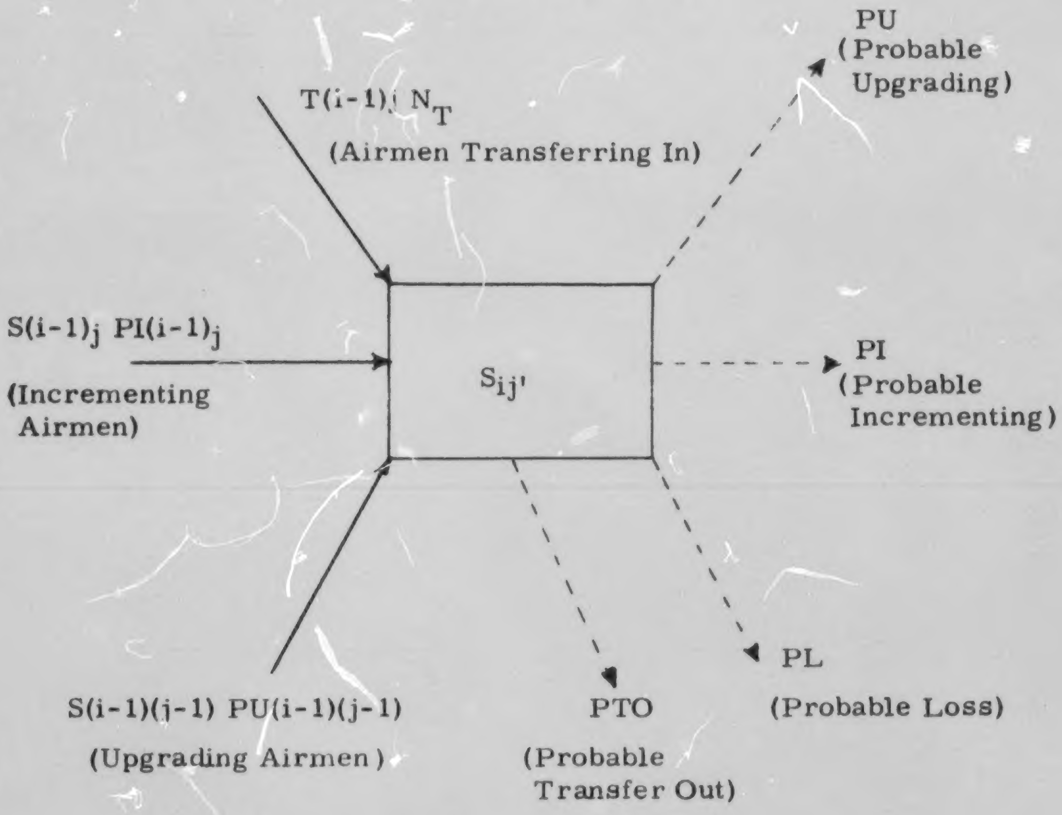
APPLICATION

- CONSTANT RECRUITMENT
- STATIONARITY



SLIDE 5

TYPICAL STATE



$T(i-1)_j N_T$
(Airmen Transferring In)

$S(i-1)_j PI(i-1)_j$
(Incrementing Airmen)

$S(i-1)(j-1) PU(i-1)(j-1)$
(Upgrading Airmen)

S_{ij}'

PU
(Probable Upgrading)

PI
(Probable Incrementing)

PL
(Probable Loss)

PTO
(Probable Transfer Out)

i = Year of Service
j = Paygrade

SLIDE 6

MARKOV MODEL

$$S'_{i,j} = S_{i-1,j}\,PI_{i-1,j} + S_{i-1,j-1}\,PU_{i-1,j-1} + T_{i,j}$$

where

$S_{i,j}$ = state (YOS=i, GRADE=j) at time t

$S'_{i,j}$ = same state at time t+1

PI = increment probability to $S'_{i,j}$

PU = upgrade probability to $S'_{i,j}$

T = number of transfers into $S'_{i,j}$

SLIDE 7

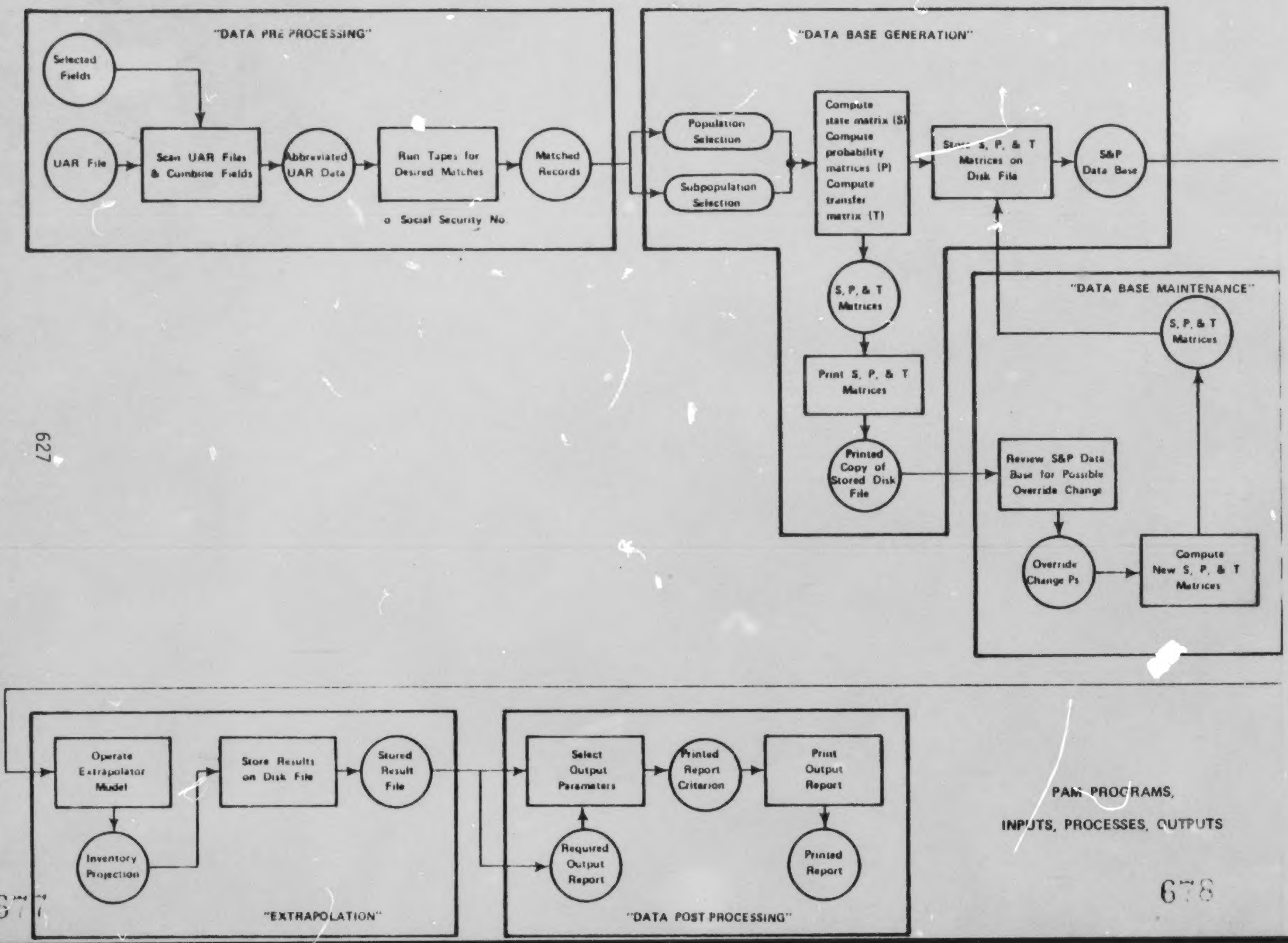
REPRESENTATIVE AIR FORCE SPECIALTY CODES

AFSC	SPECIALTY CODES
325X0	AUTOMATIC FLIGHT CONTROL SYSTEMS
325X1	INSTRUMENT SYSTEMS
328X0	AVIONIC COMMUNICATIONS
328X1	AVIONIC NAVIGATION SYSTEMS
328X4	INERTIAL AND RADAR NAVIGATION SYSTEMS
423X0	AIRCRAFT ELECTRICAL SYSTEMS
423X1	ENVIRONMENTAL SYSTEMS
423X3	FUEL SYSTEMS
423X4	PNEUDRAULIC SYSTEM
426X2	JET ENGINES
431X1E	AIRCRAFT MAINTENANCE (JET, OVER 2 ENGINES)
431X1C	AIRCRAFT MAINTENANCE (JET, 1 OR 2 ENGINES)
531X3	AIRFRAME REPAIR

SELECTED ATTRIBUTES

- PROFICIENCY PAY
- HAZARDOUS DUTY STATUS
- PRIMARY AFSC
- SECONDARY AFSC
- CONTROL AFSC
- EDUCATION ATTAINED
- YEAR GRADUATED
- TRAINING DATA
- ADMIN. TEST SCORE (%)
- ELECT. TEST SCORE (%)
- GEN. TEST SCORE (%)
- MECH. TEST SCORE (%)
- AFQT
- GRADE
- USAF SUPERVISORY EXAM RESULTS
- TOTAL ACTIVE MILITARY SERVICE
- SEX
- RACE
- BIRTH DATE
- MARITAL STATUS
- DEPENDENTS (NUMBER)
- SPECIAL EXPERIENCE IDENTIFIER
- DUTY AFSC
- UPGRADE TRAINING

SLID



"DATA PRE PROCESSING"

Selected Fields

UAR File

Scan UAR Files & Combine Fields

Abbreviated UAR Data

Run Tapes for Desired Matches

o Social Security No.

Matched Records

"DATA BASE GENERATION"

Population Selection

Subpopulation Selection

Compute state matrix (S)
Compute probability matrices (P)
Compute transfer matrix (T)

Store S, P, & T Matrices on Disk File

S&P Data Base

S, P, & T Matrices

Print S, P, & T Matrices

Printed Copy of Stored Disk File

"DATA BASE MAINTENANCE"

S, P, & T Matrices

Review S&P Data Base for Possible Override Change

Override Change Ps

Compute New S, P, & T Matrices

"EXTRAPOLATION"

Operate Extrapolator Model

Inventory Projection

Store Results on Disk File

Stored Result File

"DATA POST PROCESSING"

Select Output Parameters

Required Output Report

Printed Report Criterion

Print Output Report

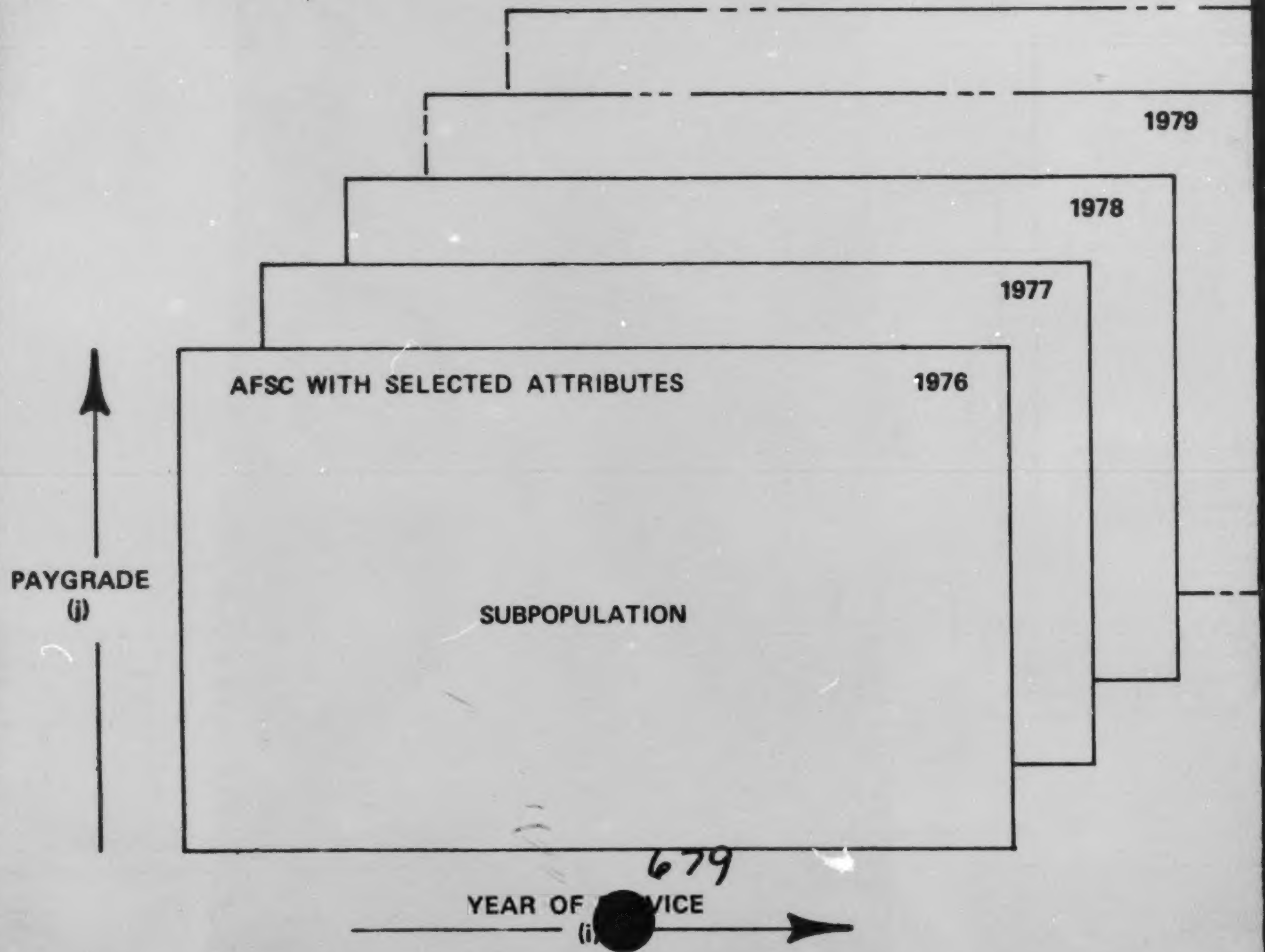
Printed Report

PAM PROGRAMS,

INPUTS, PROCESSES, OUTPUTS

SLIDE 10

SUBPOPULATION PROJECTION



1979

1978

1977

AFSC WITH SELECTED ATTRIBUTES

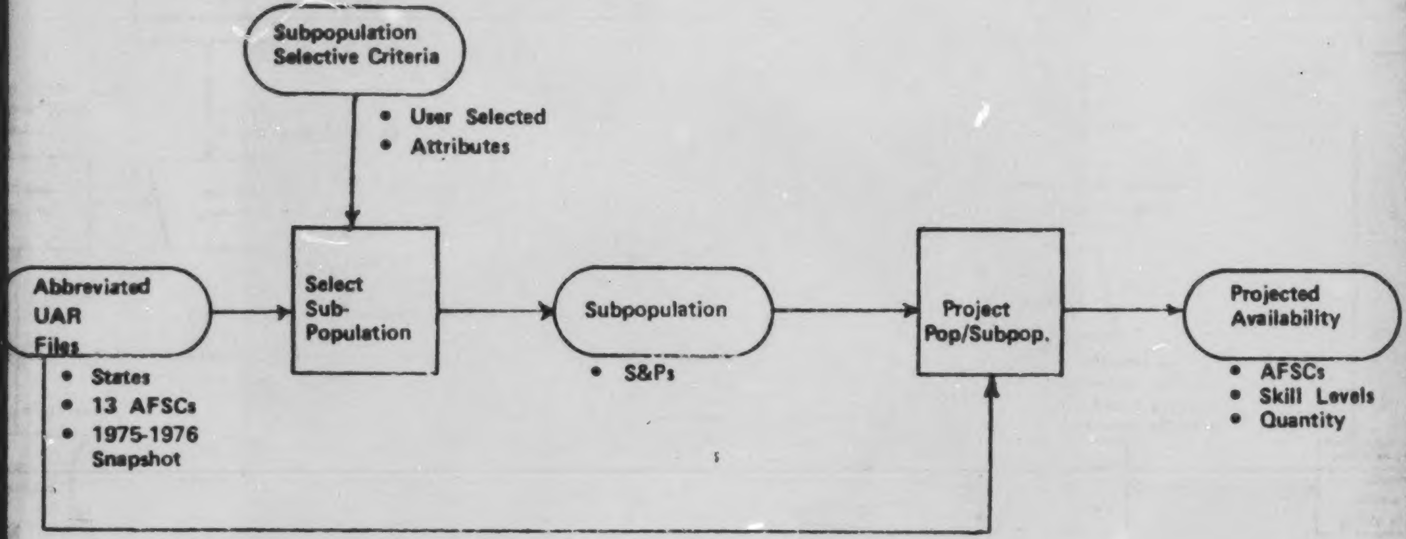
1976

SUBPOPULATION

PAYGRADE
(j)

YEAR OF [illegible]VICE
(i)

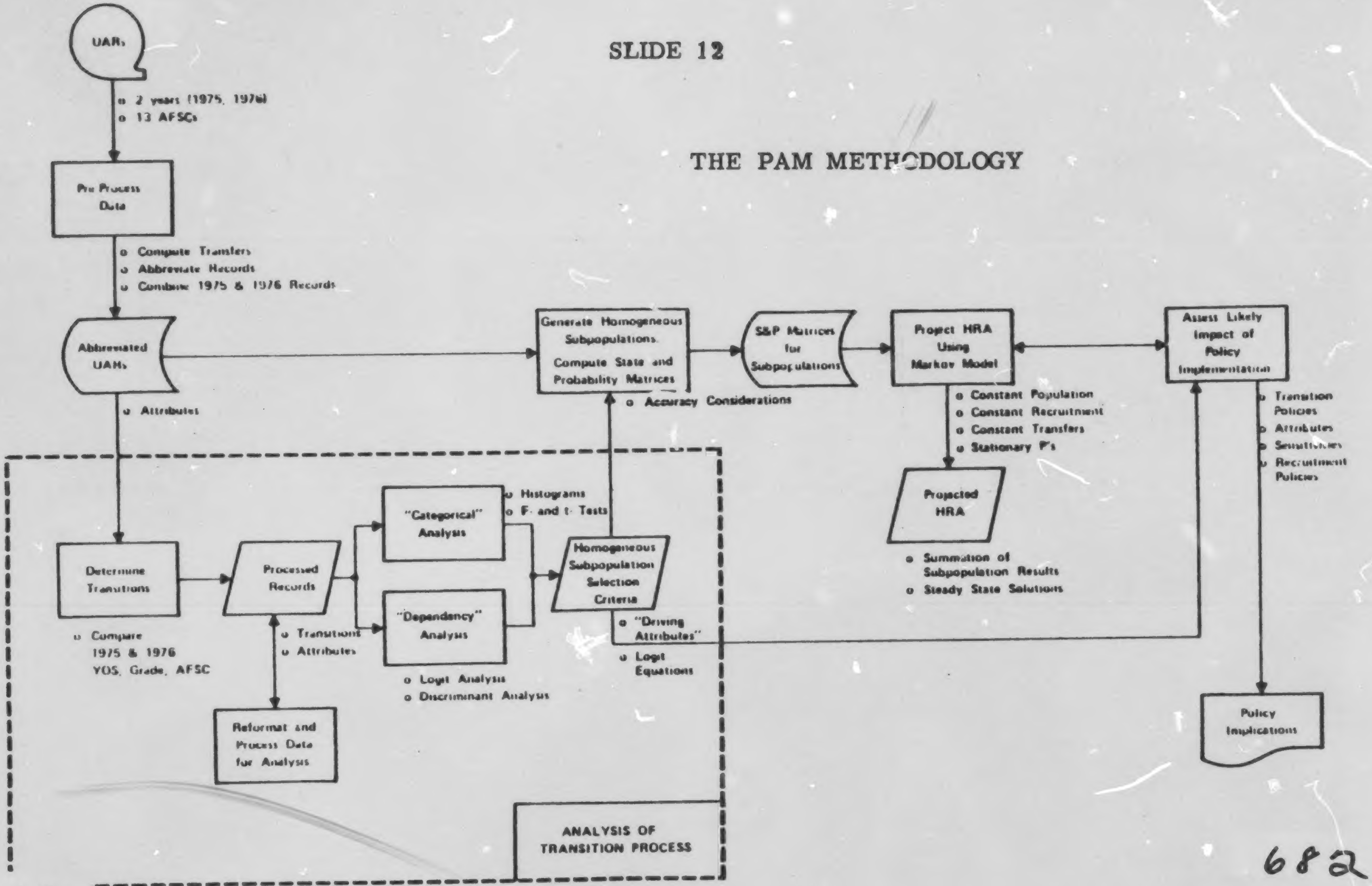
SLIDE 11


Subpopulation Selective Criteria
• User Selected
• Attributes
Abbreviated UAR Files
• States
• 13 AFSCs
• 1975-1976 Snapshot
Select Sub-Population
Subpopulation
• S&Ps
Project Pop/Subpop.
Projected Availability
• AFSCs
• Skill Levels
• Quantity

ESTABLISHMENT OF HUMAN RESOURCE AVAILABILITY

SLIDE 12

THE PAM METHODOLOGY



UARs

- 2 years (1975, 1976)
- 13 AFSCs

Pre Process Data

- Compute Transfers
- Abbreviate Records
- Combine 1975 & 1976 Records

Abbreviated UAHs

- Attributes

Generate Homogeneous Subpopulations. Compute State and Probability Matrices

- Accuracy Considerations

S&P Matrices for Subpopulations

Project HRA Using Markov Model

- Constant Population
- Constant Recruitment
- Constant Transfers
- Stationary P's

Projected HRA

- Summation of Subpopulation Results
- Steady State Solutions

Assess Likely Impact of Policy Implementation

- Transition Policies
- Attributes
- Sensitivities
- Recruitment Policies

Policy Implications

Determine Transitions

- Compare 1975 & 1976 YOS, Grade, AFSC

Processed Records

- Transitions
- Attributes

Reformat and Process Data for Analysis

"Categorical" Analysis

- Histograms
- F- and t- Tests

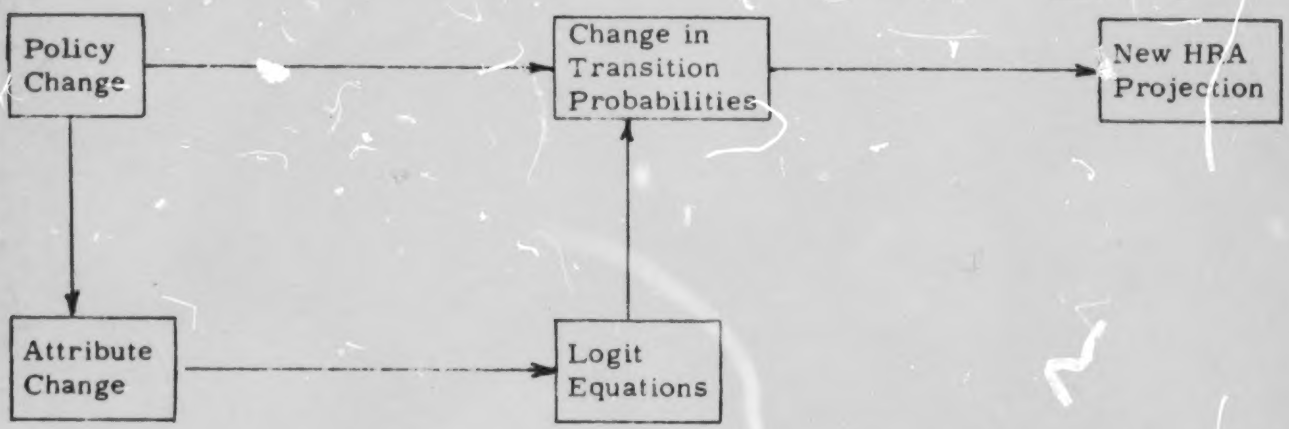
"Dependency" Analysis

- Logit Analysis
- Discriminant Analysis

Homogeneous Subpopulation Selection Criteria

- "Driving Attributes"
- Logit Equations

ANALYSIS OF TRANSITION PROCESS



Policy Change

Change in Transition Probabilities

New HRA Projection

Attribute Change

Logit Equations

POLICY ASSESSMENT

Symposium:

METHODOLOGY FOR MOBILIZATION POPULATION INVENTORY

Chairman

Jack M. Hicks

Participants

R. F. Boldt
M. A. Fischl
George V. Rux
William Graham
Lonnie D. Valentine, Jr.

Some Implications of Commercial Test Normings for Mobilization Surveys

R. F. Boldt

Educational Testing Service
Princeton, New Jersey

November 1978

Some Implications of Commercial Test Normings for Mobilization Surveys

I assume that, in an abstract way, the object of the study contemplated by this symposium can be regarded as the description of a large segment of the American population in terms of variables that are of military interest; a major problem is that we don't normally access that segment for testing. One approach is to try to estimate the statistics of interest through their relationship with variables that are included in larger, more comprehensive surveys of populations similar to the population of interest. I will, however, focus on a different alternative--that of operating one's own survey--using the development of national test norms as an example (Jackson & Schrader, 1976).

This problem is not unlike one faced by the College Entrance Examination Board, for which we (ETS) are the technical contractors. The Board owns the Scholastic Aptitude Test (SAT), a test used for admissions in a large number of American colleges and universities. It is useful to secondary school students to know where they might stand in the SAT score distributions. To supply this, the Board periodically has us undertake to find the distribution of test scores in American secondary schools. In this, we have a situation similar to that of the military. The SAT, like the ASVAB, in whose distributional statistics we might be interested, is only given to applicants; for both testing programs there is another big pool of people out there about whom we would like to know.

Actually, the Board doesn't use the SAT to construct norms; it uses a somewhat shorter test called the Preliminary Scholastic Aptitude Test (PSAT), which is extensively administered, but at somewhat lower grades than is the SAT. Use of the PSAT is feasible for several reasons. Statistical relationships between the PSAT and the SAT have been developed, and their factorial content is highly similar, as might be expected since they are built to about the same test development specifications. Also, for the schools that appear in the sample and that also regularly administer the PSAT, the job is only one of getting them to extend their testing to include the whole class (as opposed to introducing the school to a brand new administration). Since a large portion of American secondary schools regularly administer the PSAT, the reduction of effort is very substantial. Because the tests are given at somewhat different ages, conversions that include "aging the scores" have been worked out. Indeed, the procedure of developing age relationships on smaller samples and then using the relationships to transform statistics collected on younger, but available, populations is one that might well be considered for studying variables of military interest.

Next, we choose schools as the sampling unit. It's the most feasible unit for us. ETS keeps a list of American secondary schools for the College Board; it's updated monthly and if effort makes it good, it's good. It's too big a list on which to keep detailed information about any of the schools; we have just the school name and address and, for some, the principal's name. Stratification, except geographic, directly from that list is pretty well out of the question. That such a list exists is an enormous help! With a sophisticated group like this, I hesitate to say that one of the techniques for studying the population is to develop a frame, but I can hardly avoid saying it. The problem that I see for mobilization studies is that the population, as defined, is one for which it is very difficult to conceive a frame. But frame construction, at any rate, is one of the first things that I would undertake. Possibly, in sheer desperation, I would consider modifying the problem to be one where a frame can be constructed.

Next, the sample is to be drawn, and we must decide how many cases are necessary. Those of us who have been involved in quantitative research over the years know that much soul-searching goes with setting sample sizes. Usually, the survey sponsors want the samples to be small because that keeps the dollar cost down; the statisticians want the sample sizes up to get narrow confidence bands. For myself, I have arrived at the following attitude, one I can reach because of some statistical results about simultaneous confidence intervals. People don't ordinarily identify simultaneous confidence intervals with surveys of this kind, but I believe an important implication for us is there. We usually have several parameters that we want to estimate, and we would like to draw some conclusion about the set. Test norms comprise a whole series of statistics, and we don't just concern ourselves with the precision of each one at a time. Now, the finding that motivates my attitude about sample sizes is that if one sets many confidence intervals narrow enough to be useful, more cases are required than there is population! That's impossible, of course; it happens because the intervals are derived using infinite theory and the populations are finite. But the thrust of it is that whatever you do, it isn't going to be enough. Therefore, my attitude is that you should take what the traffic will bear; get all you can get if you're interested in multiple pieces of information. It's very seldom, in a social science inquiry, that I've encountered a situation in which any really credible confidence interval construction was involved; in fact, I can't remember any. Rather, there will be a certain amount of money available to do the study. What you do is to balance your resources so that you can do the best job for your sponsor and his purposes. Within that context, you get as large a sample as you can get and still get the rest of the job done--that's the state of the art. Of course, you should calculate confidence intervals for single pieces of information, and if those are too broad to be meaningful for the sample size you have, you can reasonably doubt whether the study should be performed without modification.

Therefore, by some procedure not entirely statistical, we arrived at the conclusion that there should be 200 schools! The next problem is to get their cooperation and that of their students. How (and whether) you secure that cooperation are crucial. Probably this seminar is supposed to concern technical things, but I think that careful attention to securing the cooperation of the participants is every bit as important, if not more so. In securing this cooperation, the College Board has a lot going for it. First of all, SAT scores have value, and as a consequence of participation, they can be supplied. Thus, the student has access to information that can help him forecast his position when the time comes for the grand sorting of high school students into the slots of the world, and he can get it for nothing. He gets some practice for the real thing, and the result never hurts his record. He gets [illegible] a little bit earlier in his career than he normally would, and that helps him plan his postsecondary school career strategies. The student that really doesn't have a college career in mind and had not planned to take the PSAT may, in fact, learn that there is some desirable higher education alternative open to him that he would not have known about otherwise. (Parenthetically, since we are discussing what the examinee has to gain by taking the test, let me mention that we've had very little luck with money as an incentive for students. My belief is that the amount of money needed to buy enough student cooperation to produce a good study is more than one can afford. But I do think that there has to be an incentive for examinees.)

Schools can have advantages, too. Those not usually participating in the PSAT programs will gain experience with a national testing program with which they haven't had previous experience, and will get guidance information that they didn't have. The schools that normally have their students take the SAT or PSAT will have the national norms updated, they can use them in accustomed ways and will have more complete guidance information about their student bodies. Knowing that these things are going for them, it is perfectly reasonable for the College Board to contact American secondary schools, ask them if they would be interested and willing to participate in national norms development for the SAT, and expect to get takers. I emphasize all this strongly because I think that to conduct a national study you must approach the institutions and the examinees in such a way that they have a reason to cooperate with you. In military research you can, of course, appeal to abstractions such as patriotism, and that will be effective in some cases. But when you're pursuing cooperation of a sample, you want broad acceptance. You must have appeals with nearly universal effectiveness. Therefore I think you must somehow create an approach that establishes some gain for the participants.

In any case, a letter is written to the schools and signed by the President of the College Board. It tells about the study and tries to motivate the schools to participate. At Educational Testing Service, we're very careful about who signs such letters and how they are written. Even so, we are perhaps sometimes not as careful as we ought to be. Generally, we try to find signers who are of significance to the recipients, as do you in military testing research. (I date myself by admitting it, but I have often participated in, or observed, the drafting of a letter for the Army Adjutant General's signature.) I think the source of the letter ought to be a figure who is recognizably identified with some goal of the institution whose participation you seek; the purposes given for the study need to be purposes with real appeal to the potential participants. Recruiting is a good purpose, and evaluating the relative quality of the people coming into the service is a good purpose. But they're military purposes, and may not be as directly connected with the goals of a school principal as are other uses of his time and his students' time. I can only suggest that, and can't precisely formulate, a common interest of broad appeal be established with educational associations. I think this very important part of it needs a lot of thought.

Well, we did our best. solicited the schools, and collected the data. How well did we do? In a 1974 study, the participation rate was 58.4 of the schools. We don't know why the rate was that low. With all those things going for the study, the sample obtained still wasn't that large, so we now had the problem of deciding how much of an effect the loss of schools had on the result. One way to approach this problem is to take spot surveys of the non-respondents. If you do this, it is very possible that the results will get you to your answer. Maybe that's obvious, but let me give you an example of where it worked out very neatly. In a survey of elementary schools for the National Commission for Marihuana and Drug Abuse (Boldt, Reilly, & Haberman, 1976), we were studying the types of drug education programs that were available in elementary schools. The response rate was terrible, less than 10%. Who'd believe a 4% sample? One option was to find characteristics of that 4% sample and compare them to national characteristics and find they're the same. But that isn't very convincing as a procedure, because it doesn't explain why you got the people that you got; it merely tells you some ways that they're not different. But that doesn't establish that they're the same; it just establishes that you didn't find the ways they differ. In our case, we went back to a sample of schools who didn't respond to us, and repeatedly called them until they talked to us. Their reactions were fairly monolithic. The schools thought we had originally contacted them by mistake: What, after all, did we mean asking them about a drug abuse program in elementary school? They didn't believe there was a drug problem in elementary school and felt that the survey simply didn't apply to them. In no case was there a consciously formed drug education program. The obvious point is that anything we came out with in our 4% sample is an overestimate of the magnitude of such education. We found in our data that such education occurred in very few places, that when it did occur it occurred only because officials at the next higher levels of organization wanted it, not because of pressure from the community. The survey did establish that there wasn't local, immediate pressure because of a perceived drug problem; and there wasn't much education going on. That was part of our answer, but we got it by asking the non-respondents, not by comparing information from respondents with existing statistics. Unfortunately, many times researchers do not ask the non-participants why they didn't cooperate. With schools, sometimes, it's just that the testing area isn't big enough. If you're losing schools for a reason like that you can sample or work out some other compromise.

Another technique for the non-response problem is to send a small simple questionnaire with the original letter of request and try very hard to get it completed and returned by all persons contacted. You can then perhaps get a hypothesis about why they don't participate if, in fact, they don't. Easily supplied descriptive information would be good, and I suggest that you don't leave a place to say "we are not participating because," because frankly that makes it too easy for them not to. Make the special approach to non-participants separately.

In summary, it's best if you can capitalize on some existing program. The fact that the program exists indicates that people have an interest in it, and if you can relate what you're doing to that interest you have a better chance of getting cooperation. Next, you need to establish in the minds of the possible examinees and the possible participating schools or sampling units, whatever they are, that they have a stake in this enterprise--not because they have a stake in what you do but because they have a stake in it because of what *they* do. Third, I suggest that you get all the sample that you can get, consistent with the cost and requirements of the rest of the job. Finally, make provision for collecting easily supplied data from everybody you write to in the first place, follow up those who don't cooperate, find out why, and modify your procedures, if you can, to get them in. Those aren't technical procedures (I find them remarkably unstatistical, I guess, when I think back on them), but to the extent that you need to go out and estimate national statistics, they have been crucial in our educational research.

References

Boldt, R. F., Reilly, R. R., & Haberman, P. W. A survey and assessment of drug-related problems and policies in elementary and secondary schools. In Drug use in America: Problems in perspective, Volume II of the second report of the National Commission on Marihuana and Drug Abuse, March 1973, 455-547. Also in Ronald E. Ostman (Ed.), Communication research and drug education. Beverly Hills, CA: Sage Publications, 1976. (ERIC Number ED 109 165.)

Jackson, R., & Schrader, W. B. Verbal and mathematical ability of high school juniors in 1974: A norms study of PSAT/NMSQT. RDR-76-77, No. 2. Princeton, NJ: College Entrance Examination Board, 1976.

MEASURING THE MILITARY BASE POPULATION OF THE 1980's

by

M. A. Fischl

U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

Presented in Symposium on
Methodology for Mobilization Population Inventory

20th Annual Conference of the Military Testing Association
Oklahoma City, Oklahoma
30 October - 3 November 1978

Measuring the Military Base Population of the 1980's

M. A. Fischl

U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

Abstract

Measurement of the military base population is needed to serve three general purposes. First, knowledge of the population distribution of general and specific abilities, vocational interests, and some skills, can facilitate high quality test development research. The World War II general ability measure on 12-million men has been exceedingly helpful, and it is time to update and expand it. The second purpose will be facilitation of manpower research, through obtaining population demographic, biographic, socio-economic data. In order to understand and manage the force, understanding of what is in the well seems critical. The third purpose will be facilitation of recruiting research through providing parameters of popularity and avoidance in the relevant age-group population. The paper identifies some sources of the needed data for the population measures.

Measuring the Military Base Population of the 1980's

M. A. Fischl

U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

There is a clear need to learn what the population distribution of military age young people will look like on relevant dimensions.

Samples from the population, even very large samples which join the military service over an extended period, differ very drastically on basic attributes depending on the political, economic, and defense state-of-affairs in the Unites States and the world at the time. Consider the figure below, which is the distribution of AFQT scores of all men entering the Army for the first time in two recent 12-month periods, fiscal years 1969 and 1977. In 1969 the country was at war, 1977 was a very recently completed period under all-volunteer operations. Although essentially the same in means, the dispersions are about as disparate as two distributions can get. WHAT DOES THE POPULATION FROM WHICH THESE SAMPLES WERE DRAWN REALLY LOOK LIKE?, which is the point of today's symposium.

Measurement of the military base population is needed to serve three general purposes. First, precise knowledge of the population distribution on general and specific abilities, vocational interests and perhaps some skills, can be very facilitating of high quality test development research. Knowledge of the population distribution permits research to utilize smaller samples, stratified to conform to the population distribution, than would otherwise be needed--this translates to lower costs and less disruption of operations. Knowledge of the population distribution allows for more precise estimation of psychometric relationships through enabling use of such statistics as range restriction corrections, which are dependent on such information. This is doubtless one reason why military employment test validity coefficients are invariably higher than those in private industry. The World War II general ability measure on 12-million men has been exceedingly helpful for these 30-odd years, and it is time to update and expand it to other psychological domains.

The second general purpose served by knowledge of the military base population will be facilitation of manpower research. A separate pool of information from that of the prior paragraph, but obtainable by similar methodology, consists of demographic, biographic, socio-economic variables descriptive of the population of young adults. A few years ago our office did a small feasibility examination of some sources of these data, to answer the very relevant question: "How representative is the Army?". A particular source looked at was the U.S. Office of Education (Department of HEW) National Longitudinal Study of the High School Graduating Class of 1972 (NLS). We analyzed the original data set (Spring 1972) and the first follow-up (October 1973). Some outcomes of that analysis were:

a. Longitudinal capture rate from Spring 72 to Fall 73 was 86%.

b. Of an N of approximately 20-thousand, 321--1½%--had joined the Army.

c. These 321 cases divided as follow on some dimensions of interest:

Race: 75% White, 17% Black, 8% Other

Socio-economic Status: 42% Low, 42% Middle, 14% High

Region: Northeast 18%, North Central 27%, South 35%, West 20%.

High School Activities: Athletics 56%, Vocational Educational 21%, Hobby Clubs 22%, Drama, Debating, Music 25%.

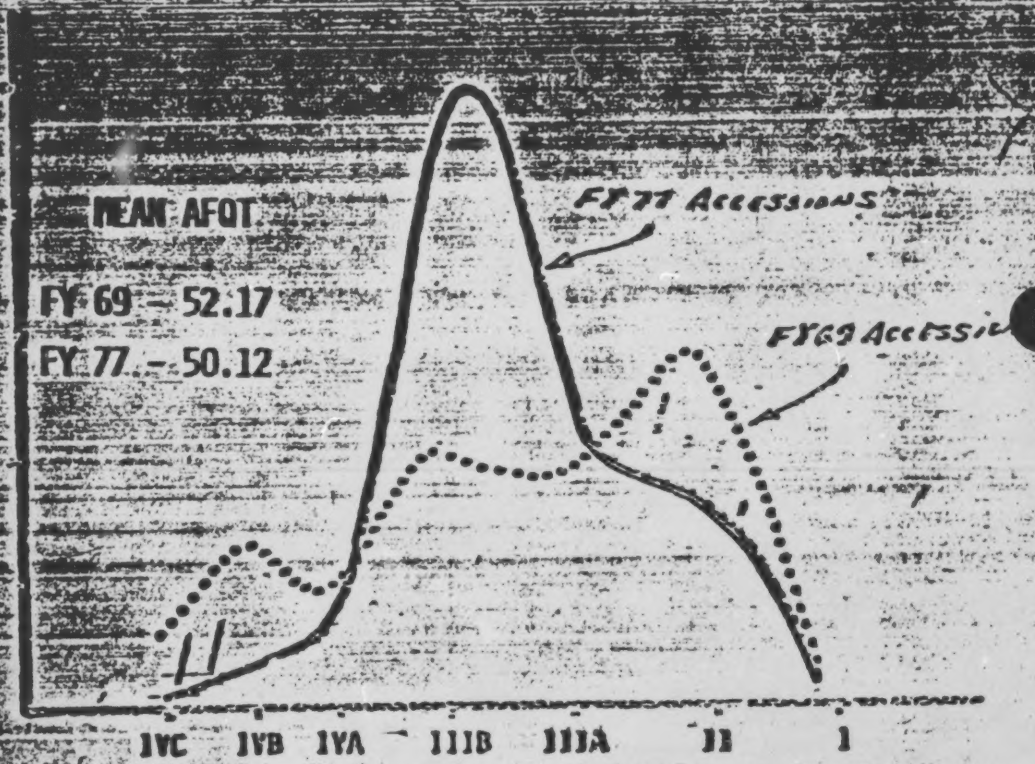
d. We did not review all dimensions.

The point here is not to report on the NLS but to indicate that population data on these types of variables can be helpful to manpower research. To understand and manage the force composition, understanding of what is in the well seems critical.

The third purpose we see facilitated by the mobilization population data set will be recruiting research, providing precise quantification of domains of military service perceived to have positive and negative valences, and providing stable benchmarks for evaluation of changes in Service recruiting policies.

How may these ends be served, yielding a census of young people in terms of the types of variables I described? Clearly we will need to splice together data from several sources, coupled with some special-purpose empirical data collection. It seems apparent that the High School ASVAB Testing Program can be of great value, and splicing it to the production AFEES Testing Program seems a natural; reliance on and tying into subsequent NLS follow-ups would seem profitable; lessons to be learned from Project TALENT should be sought; there are numerous ad hoc and continuing Department of Labor, Department of Defense (e.g., Gilbert Youth Survey), Bureau of Labor Statistics and Bureau of the Census surveys to be examined. I've just skimmed the surface and the obvious. The other participants today will tell us of prior experiences in this type of endeavor inside and outside the Department of Defense and I hope illuminate a way that we can cooperatively inventory the military base population of the 1980's.

[illegible]


MEAN AFQT
FY 69 – 52.17
FY 77 – 50.12
FY 77 ACCESSIONS
FY69 ACCESSI
IVC
IVB
IVA
IIIB
IIIA
II
I

DEVELOPMENT OF A MOBILIZATION POPULATION INVENTORY USING EXISTING ASVAB DATA BANKS

BY: GEORGE V. RUX
WILLIAM GRAHAM

DEVELOPMENT OF A MOBILIZATION POPULATION INVENTORY USING EXISTING ASVAB DATA BANKS

The Military Enlistment Processing Command (MEPCOM) annually tests approximately 1.9 million individuals with the Armed Service Vocational Aptitude Battery (ASVAB). Approximately 800,000 are applicants for enlistment in the Armed Services and are tested with the ASVAB Form 6 or 7. The applicants are tested at one of the 66 Armed Forces Examining and Entrance Stations (AFEES) or at one of over 1,000 testing locations run by the AFEES. Additionally, there are approximately 1.1 million high school students who are annually administered the ASVAB Form 5 in their own high schools. Results of this high school version are then used to provide a prescreened list of mentally qualified prospects for enlistment.

The mobilization population can be defined as the set of 18-24 year old, American citizens. Presumably this subset of the American population would be subject to the draft during a national emergency. Naturally, it would be desirable to test a large unbiased sample of the mobilization population to develop a new mobilization base, but the cost would probably be prohibitive. Accordingly, we may have to sacrifice some theoretical purity in the face of economic constraints. Nevertheless, the possibilities appear bright for tailoring existing data to develop a new mobilization base which can be more accurate and of greater utility than the one currently in use. The data base that appears most feasible for use in modeling the mobilization population is the high school sample. There are two reasons for this selection: low cost and low pre-test bias.

698

First, the sheer magnitude of the number of students tested, combined with the demographic data that is coded by the students on their answer cards, allows for the instant computerized analysis of extremely large samples. Additionally, the results obtained on the ASVAB 5 are relatively unbiased by illegal pretest assistance. Unfortunately, applicants for enlistment are frequently provided pretest information regarding items on the ASVAB 6 or 7. This behavior (commonly called compromise) is a major factor inflating the scores of applicants on the Armed Forces Qualification Test (AFQT), which is the qualification portion of the ASVAB. In contrast to the ASVAB 6/7 production testing, the effect of test compromise in the high school testing program appears to be negligable. Only 8% of those taking the test initially indicate a desire to enter the military. Additionally, those who intend to obtain illegal assistance probably opt for immediate testing on the ASVAB 6/7 rather than waiting for their high school to schedule the ASVAB 5. Figures 1 and 2 are examples of ASVAB 5 vs ASVAB 6/7 percentile plots for male and female applicants, respectively, who were administered both tests. The ASVAB 5 sample appears to be closer to a "normal curve" than the curve depicting the same sample of individuals who took the ASVAB 6/7 version of ASVAB.

There are two basic problems with estimating the performance characteristics of our referenced population, no matter what sample we take:

(1) The motivation of the individuals within the sample.

(2) the appropriateness of the sample.

The motivation factor, unfortunately, has pervasive effects with our current method of norming. During the draft era, a signficant portion of selective service registrants would intentionally fail the AFQT in the hope they would not be inducted. Now the problem is reversed. A significant portion of applicants has received some form of unauthorized testing assistance so that they will be found mentally qualified for the service and job of their choice when, in fact, they are unqualified. The current method of norming new tests requires stratification of a sample of applicants in the AFEES by AFQT. Unfortunately, since the existing AFQT is compromised, the new items appear harder because the stratified sample makes applicants appear more capable then they really are. Our norms are continually degenerated each time this stratification process takes place because the effects are cumulative. In essence, we are stratifying the population to insure the sample is unbiased but we are not compensating for the overriding effect of test compromise.

While it is difficult to quantify the extent of compromise, its effect can be demonstrated using a verification composite first proposed by Sims (reference 1). Figures 3 through 6 are percentile plots based on all applicants tested in all AFEES from Apr - Jun 78. The qualification composite (AFQT) is compared against a composite based on other non-AFQT tests within the ASVAB. For ease of discussing, this verification composite is called the "pseudo" AFQT. The difference in these curves reflects different rates of compromise by service. Unfortunately there is no readily available "clean" sample of applicants upon which to measure the true extent of compromise in the current AFQT versions.

700

There are additional indications of test compromise. MEPCOM recently instituted a statistical procedure to identify those applicants who had inconsistently high qualification scores, so that they might be retested. A "5% screening table" was developed using the cross tabulated test scores of a sample of over 40,000 applicants for enlistment. The screening table is an internal consistency check: the performance on individual tests within the AFQT is compared to the performance on highly correlated tests elsewhere in the battery. The tables are statistically designed to screen 5% of all those applicants whose test score comparisons appear most aberrant.

Using this screening table on a sample from the AFEES, it is readily apparent the AFQT is compromised, especially the Word Knowledge test. Table 1 shows a comparison of this nature where a "clean" sample of 3134 by USMC recruits were tested in early 1976 when the ASVAB 6/7 was initially introduced. Ideally, what is needed is some form of internal consitency check of item distractors to eliminate applicants who were "coached" from the sample while still retaining an unbiased sample of the mobilization population. In this fashion, those individuals who purposely failed most of the easier distractors could be detected as inconsistently low. By the same method, those who consistently failed the difficult distractors yet who scored high enough to qualify could be identifed as inconsistently high. Until internal consistency checks are instituted, the current practice of standardizing tests in the AFEES using the existing qualification test should be used only as an interim solution.

A recent preliminary report of the descriptive statistics for the ASVAB 5 (reference 2) indicates that motivation is not a problem. Few students intentionally do poorly on the ASVAB 5. The students taking the ASVAB 5 are not, however, a random selection of all high school students within the nation. For most students the decision to take the test is voluntary since the DOD does not require the test. In our recent analysis, 30% of the students tested indicated a desire to attend a four year college program. In addition, a greater percentage of students from the south elect to take the ASVAB than would be expected by examining population density statistics. Nevertheless, sample bias can be overcome. This school year (77-78), for the first time, data is being differentiated on the basis of which testing sessions are mandatory and which are optional. (Mandatory sessions are those for which the high school counselors have elected to test all students within a given grade). In essense, it is now possible to obtain representative statistics on those students who previously chose not to take the ASVAB because of a predominate interest in attending college. We can now compensate for sample bias by statistically selecting test data from mandatory testing sessions whose aggregate population reflects the demographic characteristics of the nation in terms of the following characteristics:

(1) Population density (by zip code region)

(2) Race

(3) Sex

(4) Plans after graduation

Initial norms based on high school testing results appear promising. Referring to the information in Figure 7, we have a comparison of a norm based on a random sample of ASVAB 5 results against the norms actually used for enlistment qualification. By demographically stratifying the sample, the curve may be shifted to the right somewhat but it is apparent that this student sample will better describe the mobilization population.

It should be clear from the forms of the two curves, a sample of over 200,000 students will describe a smoother and more representative curve then a sample of preselected recruits. At the present time MEPCOM is testing one out of every six seniors in the nation. By the use of prudent statistical sampling one can have, at reasonable cost, a large data base that is demographically stratified to reflect an unbiased sample of the nation's mobilization population.

REFERENCES

1. Center for Naval Analyses, (CNA) 78-3052, "A Review of MEPCOM Verification Retest Proposal", by William H. Sims, Unclassified, 27 Apr 78.

2. Draft Technical Report for Descriptive Statics for ASVAB 5, DAKF15-77-C-0188, submitted by Charles T. Kenny, Ph.D., Project Manager, Unclassified, School Year 1976-1977.

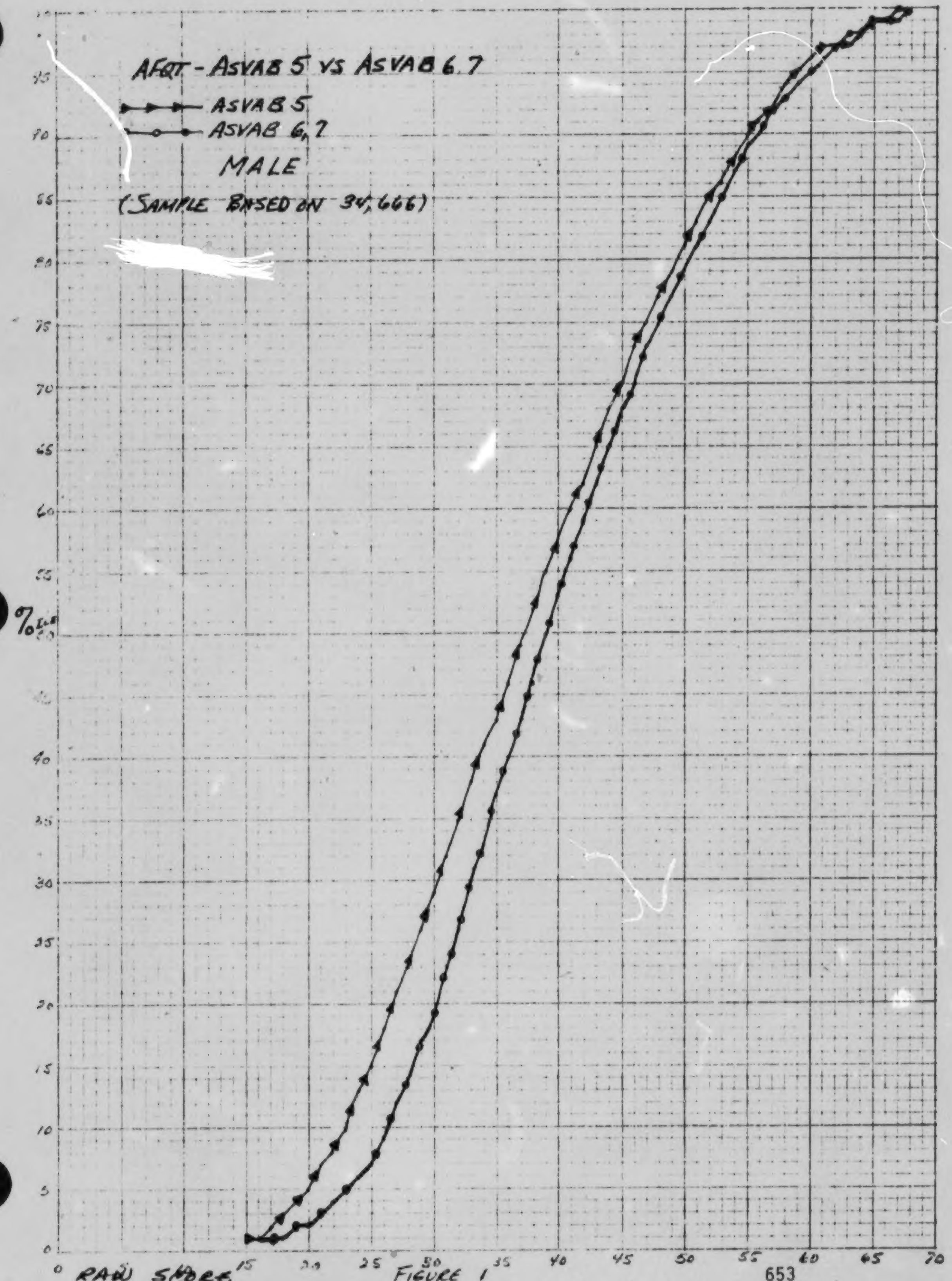

AFQT - ASVAB 5 VS ASVAB 6,7
ASVAB 5
ASVAB 6,7
MALE
(SAMPLE BASED ON 34,666)
%ile
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
15
20
25
30
35
40
45
50
55
60
65
70
RAW SCORE

FIGURE 1

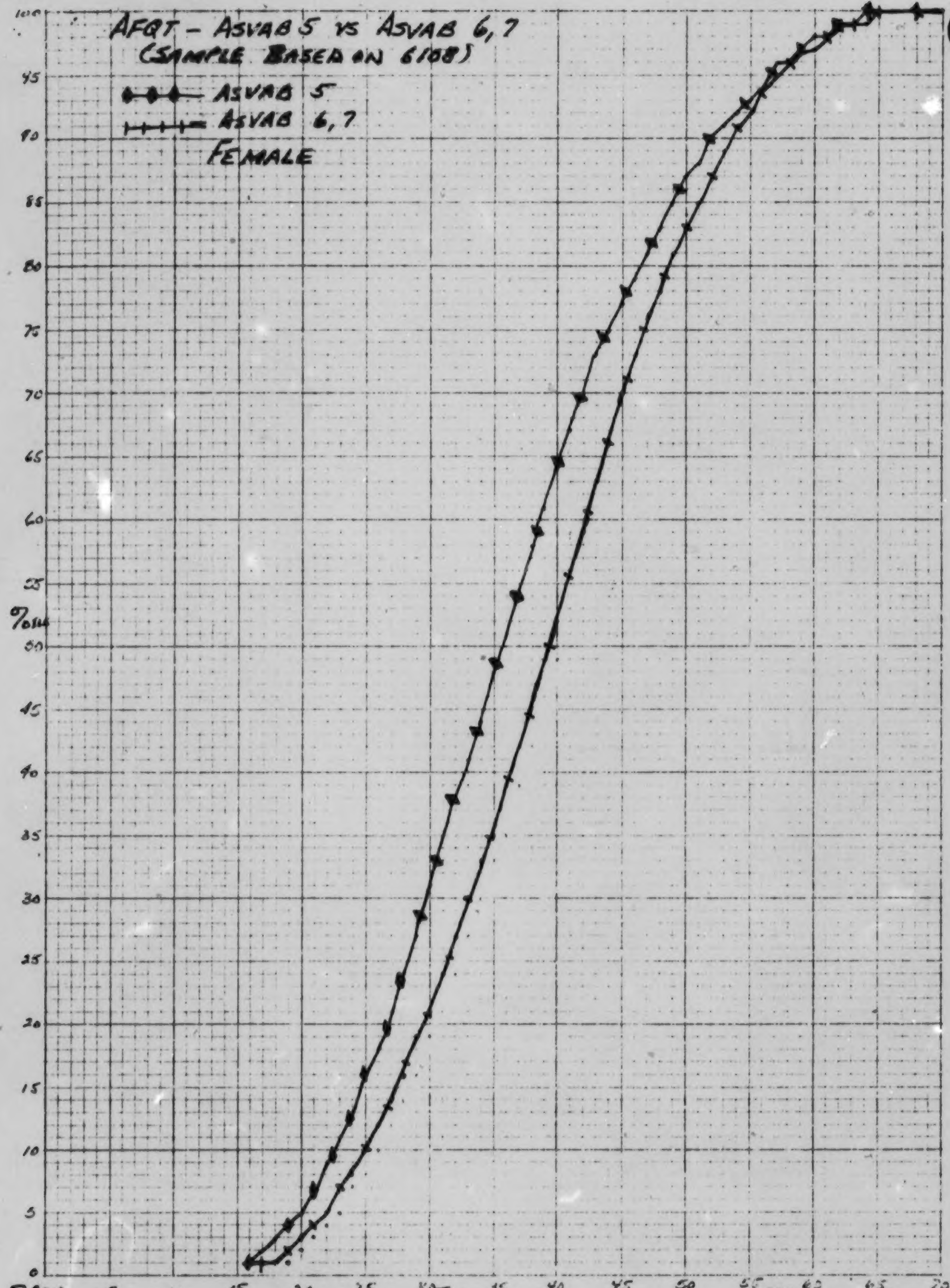

AFQT - ASVAB 5 vs ASVAB 6,7
(SAMPLE BASED ON 6108)
ASVAB 5
ASVAB 6,7
FEMALE
%ile
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
RAW.
5
10
15
20
25
30
35
40
45
50
55
60
65
70

FIGURE 2

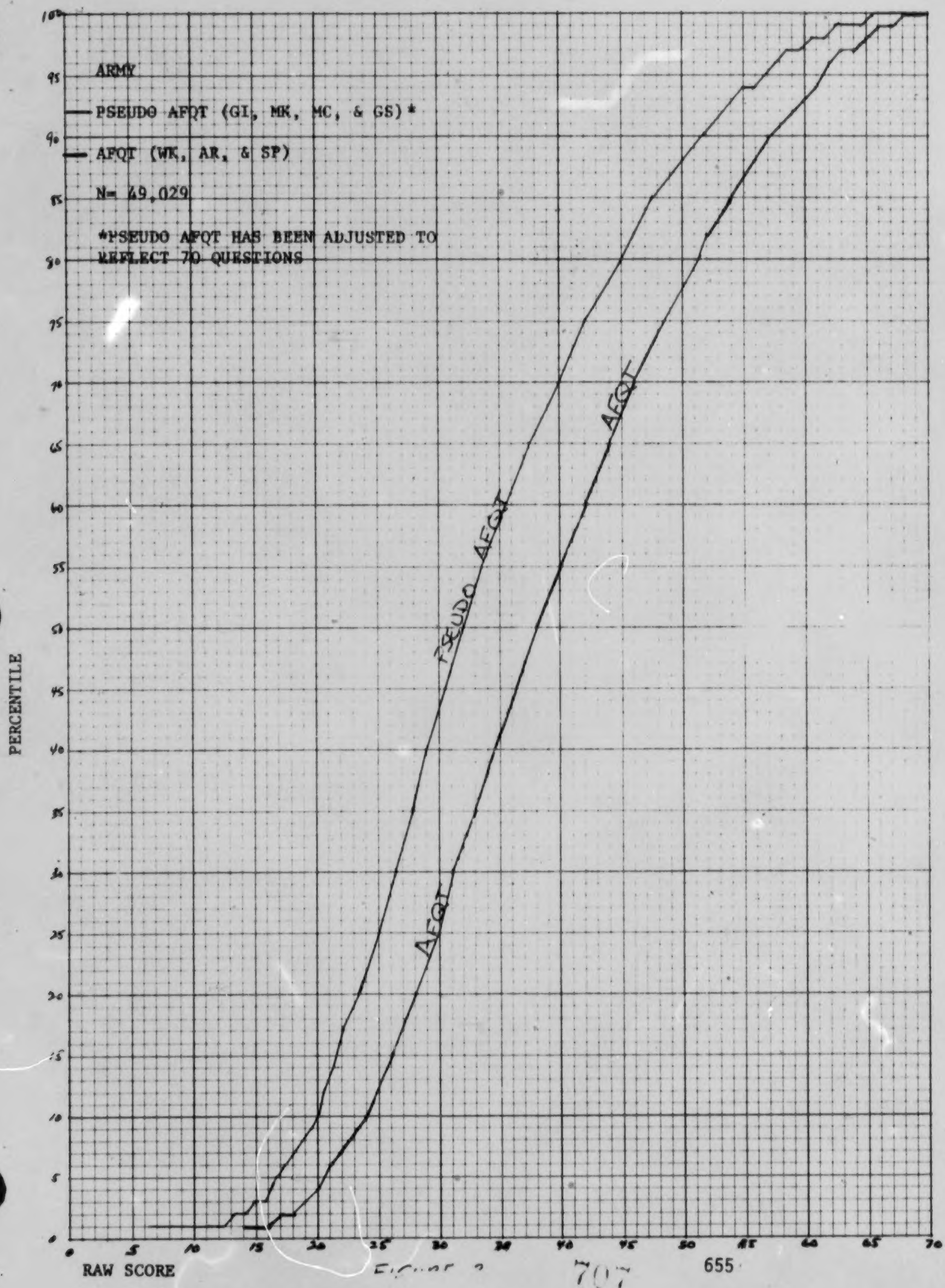

ARMY
PSEUDO AFQT (GI, MK, MC, & GS)*
AFQT (WK, AR, & SP)
N= 49,029
*PSEUDO AFQT HAS BEEN ADJUSTED TO REFLECT 70 QUESTIONS
PSEUDO AFQT
AFQT
AFQT
PERCENTILE
RAW SCORE
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

FIGURE 3

707

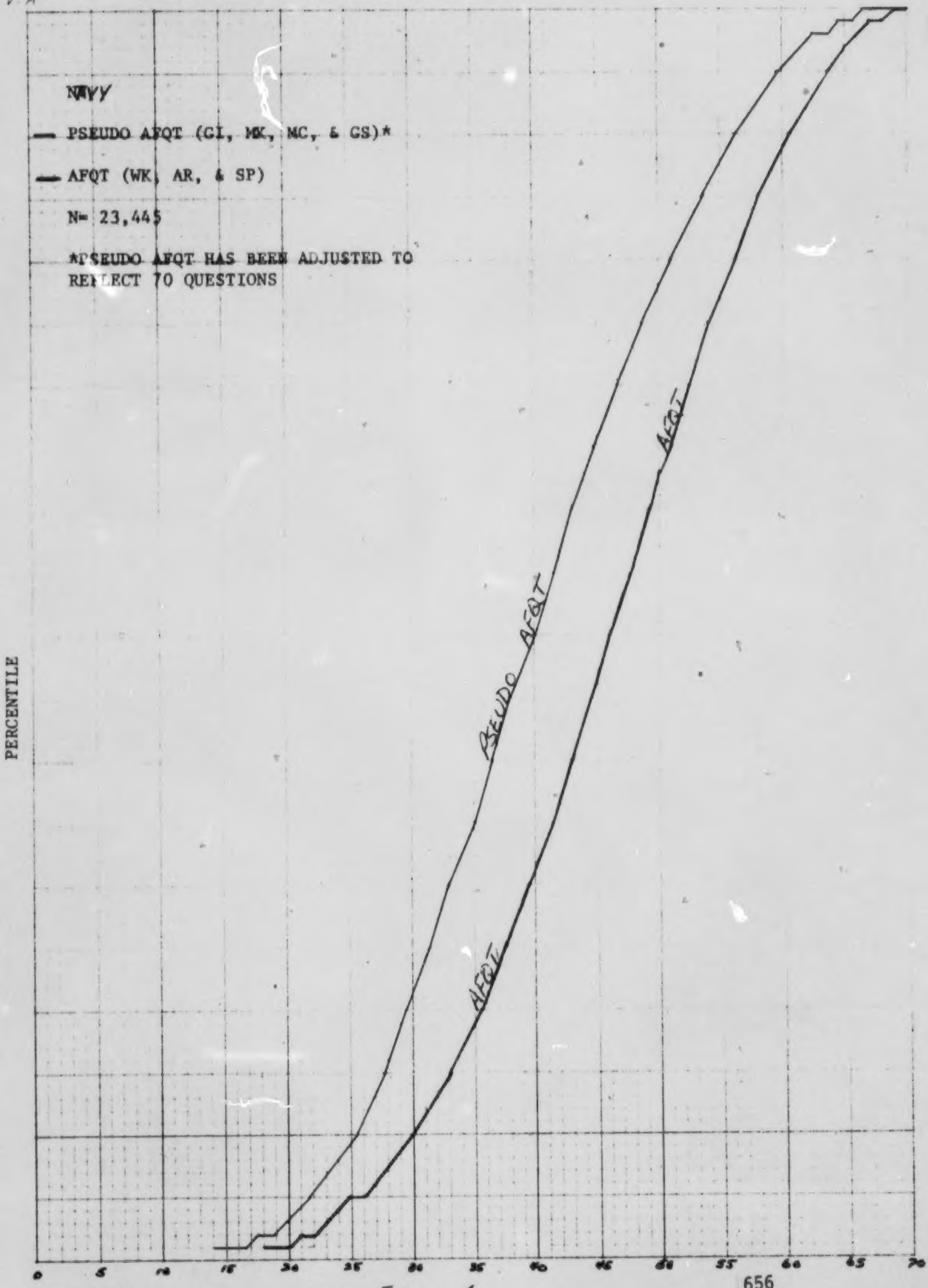

NAVY
PSEUDO AFQT (GI, MK, MC, & GS)*
AFQT (WK, AR, & SP)
N= 23,445
*PSEUDO AFQT HAS BEEN ADJUSTED TO REFLECT 70 QUESTIONS
PSEUDO AFQT
AFQT
AFQT
PERCENTILE
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
RAW SCORE

FIGURE 4

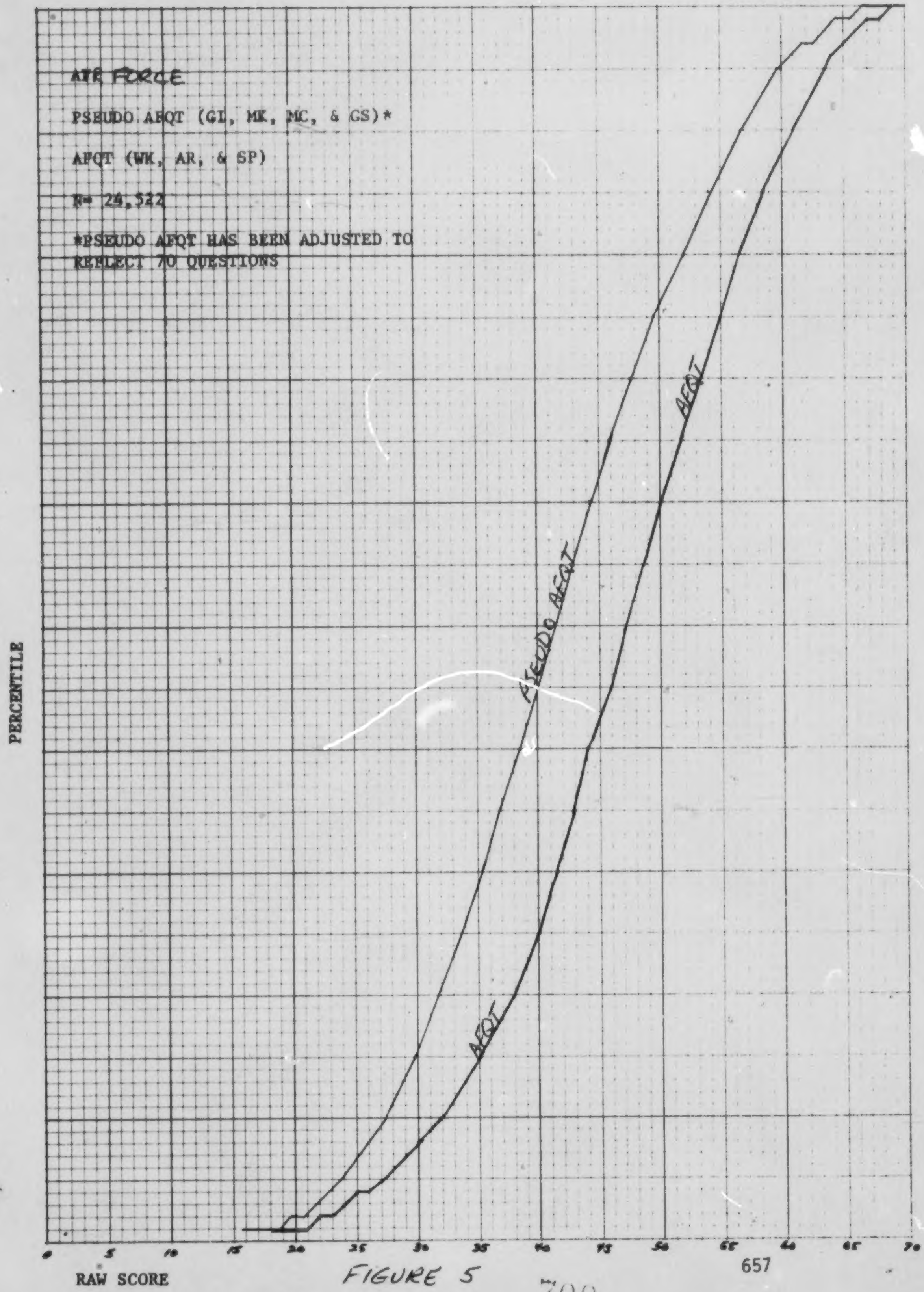

AIR FORCE
PSEUDO AFQT (GI, MK, MC, & GS)*
AFQT (WK, AR, & SP)
N= 24,522
*PSEUDO AFQT HAS BEEN ADJUSTED TO REFLECT 70 QUESTIONS
PSEUDO AFQT
AFQT
AFQT
PERCENTILE
RAW SCORE
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70

FIGURE 5

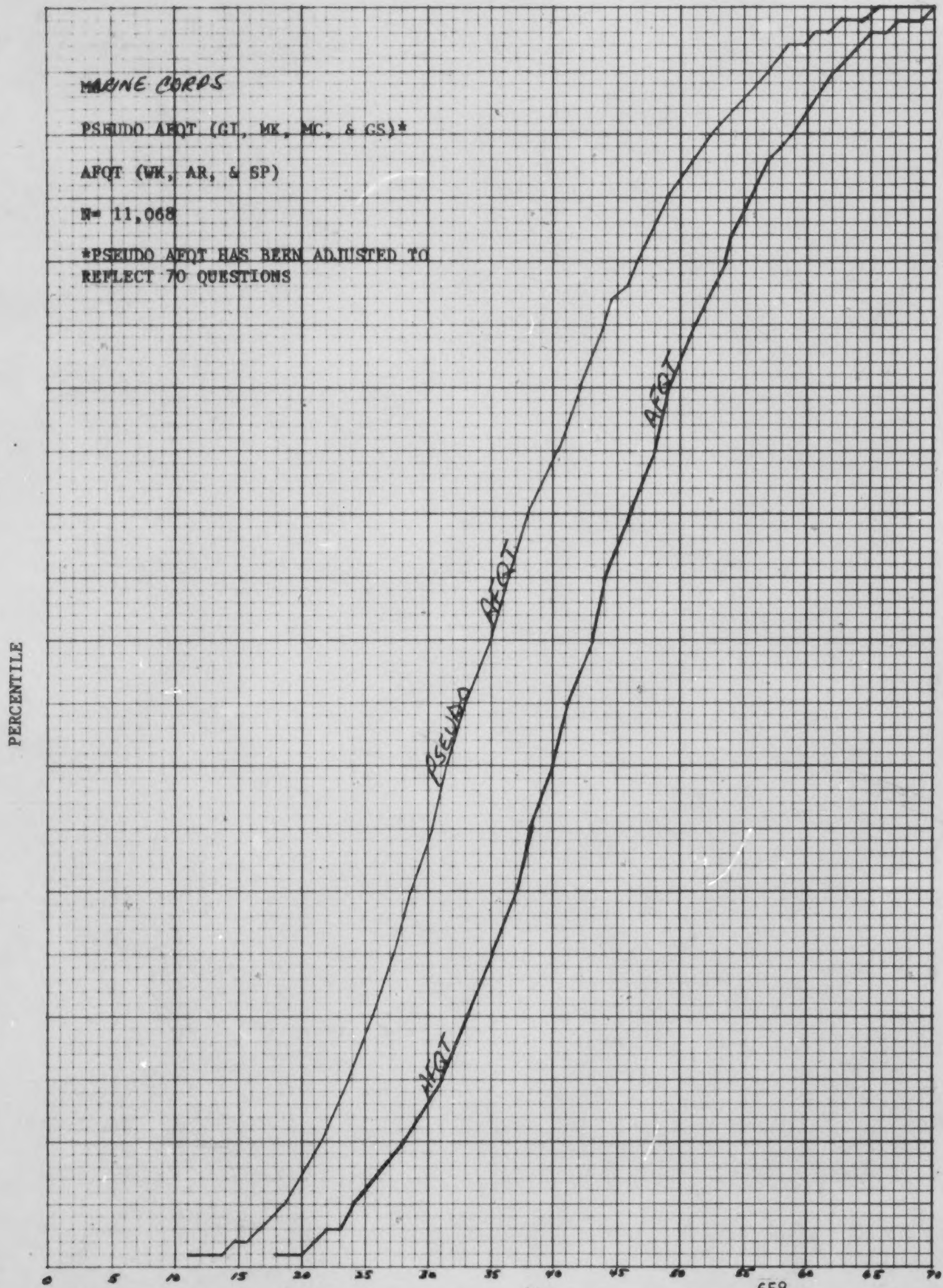

MARINE CORPS
PSEUDO AFQT (GI, WK, MC, & GS)*
AFQT (WK, AR, & SP)
N= 11,068
*PSEUDO AFQT HAS BEEN ADJUSTED TO
REFLECT 70 QUESTIONS
PSEUDO AFQT
AFQT
AFQT
PERCENTILE
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
RAW SCORE

FIGURE 6

TABLE 1

SCREENING EFFECT ON
CLEAN AND OPERATIONAL SAMPLES OF USMC RECRUITS

ASVAB TEST	CLEAN (3134 RECRUITS) (1) #FAILED	CLEAN (3134 RECRUITS) (2) %FAILED	OPERATIONAL (4896 RECRUITS)[A] (3) #FAILED	OPERATIONAL (4896 RECRUITS)[A] (4) %FAILED
WK	30	0.96%	322	6.58%
AR	16	0.51%	28	0.57%
SP	11	0.35%	19	0.39%

[A] April 1978 Marine Corps Applicants.

CUMULATIVE PERCENTILE CURVES
FOR
HIGH SCHOOL STUDENTS VS NORMS
FOR
ASVAB 5

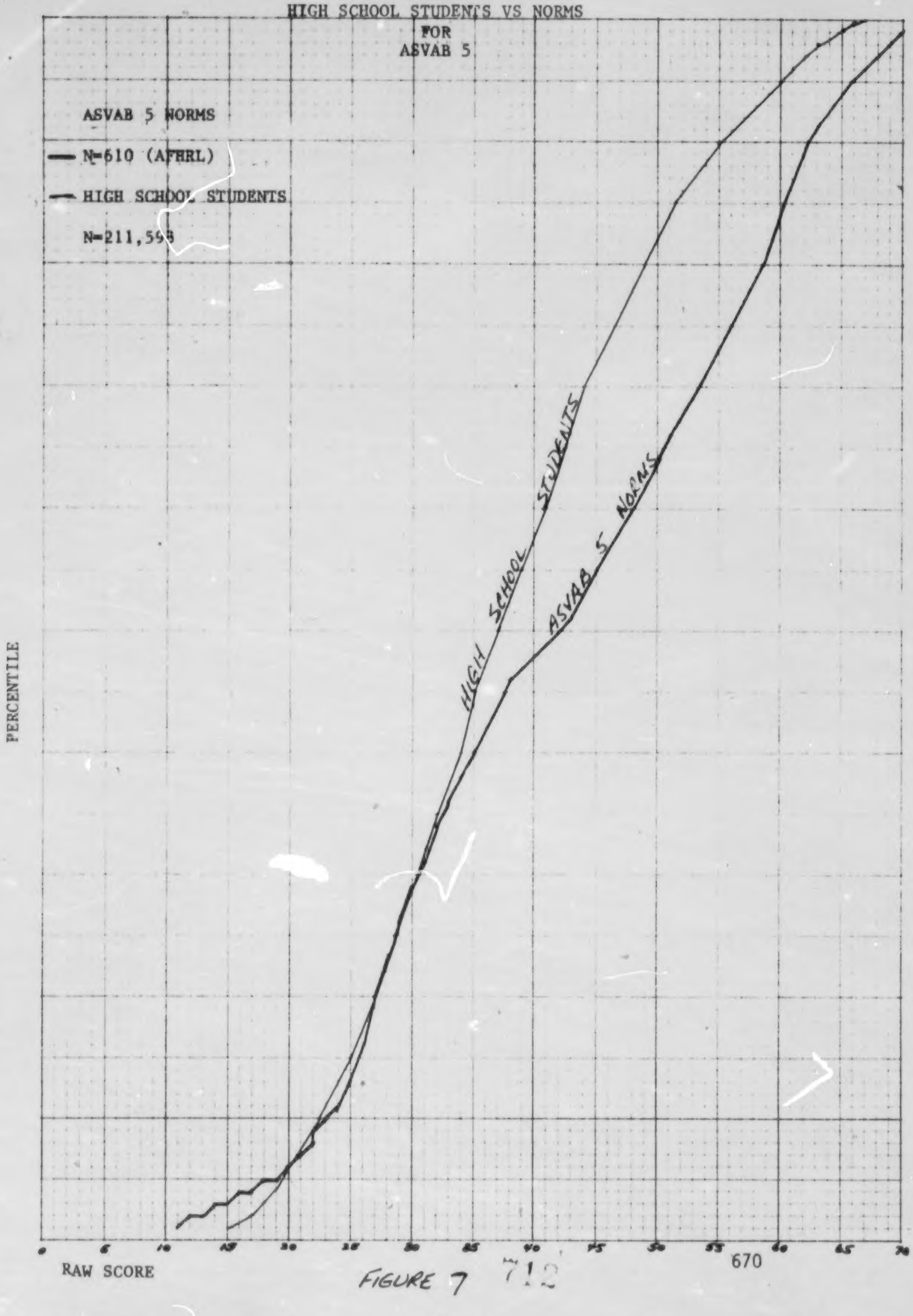


FIGURE 7

Air Force Experience with PROJECT TALENT

Lonnie D. Valentine, Jr.
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Over the years, one of the chief means used by test constructors for maintaining stable normative standards (or score meaning) from one form of a test to the next has been through equipercentile conversion procedures in which each form of the test is calibrated such that its distribution matches that of a "reference" test which has been calibrated for the target population.

In the case of the Armed Services, the most frequently used enlisted test standarization reference measure has been the Armed Forces Qualification Test (AFQT). The predecessor test to AFQT was standardized during World War II on a very large sample, representative of service-age young men.

Since that time, virtually all service selection and classification test batteries for enlisted personnel have been calibrated against those World War II standards through AFQT. This has generally been true regardless of whether the score being standardized was intended as an alternate form of AFQT.

For its officer test programs, the Air Force, in 1954-55, adapted as its standards reference, distribution of ability among Air Force Academy (AFA) applicants. The assumption was that academy applicants were a select group of young men who could well define the standard against which officer applicants were to be compared. Moreover, because AFA is a prestige program, ability levels of AFA applicants were assumed to be fairly constant from one year's applicants to the next. Prior to that time, officer standards were calibrated in terms of performance of World War II aircrew program applicants.

Academy applicants did, in fact, prove to be a select group of young men, but over time the nature of their selectivity was such that they became inadequately representative of the "target" pool for other officer programs. Their performance on both verbal and quantitative ability measures was initially equivalent/above average for 18-year-olds, but, over the first few academy classes, the applicants became increasingly self-selected on quantitative abilities while their levels of verbal ability held constant. Thus, if one established norms for successive AFOQT's directly on raw score distributions among academy applicants to successive classes, the verbal norms would have held fairly constant within the broader officer target population, but quantitative norms would be badly biased (i.e., relatively higher raw scores would be associated with moderate or low converted scores).

If one considered the sum of College Entrance Exam Board/Verbal and Quantitative, indicators of general ability, as the reference measures, one would have a circumstance in which the use of these reference measures would result in "easy" verbal standards and "difficult" quantitative standards.

This brings us to an important principle with respect to equipercentile norming procedures, specifically: the stronger the relationship between the normative reference measure and the measure to be standardized, the higher the probability that the new norms will be unaffected by atypical sampling and uncontrolled variables.

Air Force began reviewing its test standardization procedures in light of this principle and concluded that it would be highly desirable to use a different reference measure for each test or composite to be normed, with the reference measures selected to correlate as highly as possible with the score being normed; if an Airman Qualifying Examination were being normed, it would be desirable to use separate Mechanical, Administrative, General, and Electronics AI reference measures. Ideally, these reference measures would be parallel forms of their counterpart in the new battery and would correlate with it or about the order of reliability.

What was needed was a large data base--lots of subjects, and a broad content spectrum of measurements from which appropriate composites could be developed, normed, and then used as benchmarks.

At about this time, the American Institutes for Research (AIR) was starting PROJECT TALENT, a national aptitude census study. Contractual arrangements were made with AIR for linkage of Air Force tests to the PROJECT TALENT data base such that a composite of TALENT measures might be developed as a normative reference for each separate Air Force selection and/or classification test or composite. The study through which these reference composites were developed is detailed in an Air Force technical report by Dailey, Shaycroft, and Orr (1962).

The TALENT Battery was administered to approximately 3,300 basic airmen, yielding about 2,500 complete cases, stratified by Armed Forces Qualification Test (AFQT) deciles in the centile range 21-100. The Air Force provided records of subtest and composite scores for the AFQT, the Air Force Officer Qualifying Test (AFOQT), and the Airman Qualifying Exam (AQE) for each airman in the sample.

For the data analysis, the total sample was randomly divided into two approximately equal subsamples, designated Subsample A and Subsample B. Much of the data analysis was done separately for the two subsamples.

In order to pick the best predictive composite for each of the Air Force variables, multiple regression analyses were run with each one of the Air Force variables, in turn, as criterion, and with 74 of the TALENT Battery test scores as predictors.

On the basis of these analyses, sets of predictor variables were selected from TALENT Battery reference composites for the Air Force criterion variables. One restriction on this selection was to limit the total testing time for the tests predicting AFOQT to 4 hours and for those predicting AQE to 2 hours. Prediction weights were expressed as integers, roughly proportional to the Sample A and Sample B average of the raw score regression weights obtained by a stepwise regression procedure. Typically, these TALENT based composites correlated about .8 with the composites they were designed to predict on cross-application to other samples.

This study allowed for estimation of the distribution of 18-year-old performance on the various Air Force tests and provided constant highly correlated reference measures for norming future revisions of the Air Force tests. These reference composites were used for a number of years in Air Force test norming studies. Air Force experience with these reference composites leads to a few recommendations I would like to pass on with respect to an appropriate mobilization base study.

(1) The test battery for such a study should be quite broad both in content and difficulty range. In our experience, the TALENT Battery was quite adequate for enlisted reference composites; however, for officer test norming studies, more "top" on the battery would have been desirable. It would seem entirely appropriate to define the mobilization base in terms of the manpower pool available both for enlisted and officer specialties; thus, measures in the battery should accommodate a broad spectrum of ability. Broad content coverage is needed both to permit initial development of highly relevant reference composites and to permit later development of new reference composites as tests and test programs change.

(2) The study's sampling plan should include adequately large representation of the potential officer pool. One product of the study should be standards against which officer and aircrew tests can be normed. It would be desirable to code data on participants in the study such that specific subpopulations of possible service interest may be identified and used as a standards reference.

(3) The study should provide a standards reference data base which may be extended as service tests change over time. While established reference composites should be retained in the data base, specific subtest data should also be retained in an easily accessible form. Whenever new service tests are developed, it should be possible to relate the new test to the subtests of the mobilization base battery, to develop a reference composite for the new test, and to establish reference conversion standards in the mobilization base file.

Basically, a mobilization survey should have broad content coverage, encompass a broad ability range, and should be easily exercised to form highly correlated reference standards for both current and future service tests.

REFERENCE

Dailey, J.T., Shaycoft, M.R., & Orr, D.B. Calibration of Air Force selection tests to PROJECT TALENT norms (PRL-TDR-62-6). Lackland AFB TX: 6570th Personnel Research Laboratory, May 1962.

20TH ANNUAL CONFERENCE

OF THE

MILITARY TESTING ASSOCIATION

PROCEEDINGS

COORDINATED BY
UNITED STATES
COAST GUARD INSTITUTE

VOLUME II

OKLAHOMA CITY, OKLAHOMA
30 OCTOBER – 3 NOVEMBER 1978

717

PROCEEDINGS

20th Annual Conference
of the
Military Testing Association

Coordinated By:

U. S. COAST GUARD INSTITUTE
P. O. Substation 18
Oklahoma City, Oklahoma 73169

HILTON INN WEST
Oklahoma City, Oklahoma
30 October - 3 November 1978

Volume II

SECTION 9

VALIDATION-PREDICTION

THE IMPACT OF VALID SELECTION PROCEDURES ON WORKFORCE PRODUCTIVITY

Frank L. Schmidt
U.S. Civil Service Commission
and
George Washington University

John E. Hunter
Michigan State University

Robert C. McKenzie
U.S. Civil Service Commission

Tressie W. Muldrow
U.S. Civil Service Commission

Running head: The Impact of Valid

The opinions expressed herein are those of the authors and do not necessarily reflect official policy of the U.S. Civil Service Commission. Requests for reprints should be sent to Frank L. Schmidt, Personnel Research and Development Center, U.S. Civil Service Commission, 1900 E Street, N.W., Washington, D.C. 20415.

The Impact of Valid Selection Procedures on Workforce Productivity

Abstract

This study reports evidence showing that the impact of valid selection procedures on workforce productivity is much greater than personnel psychologists have typically believed. The Brogden-Cronbach decision theoretic models of selection utility are presented and explained. The three major reasons for the failure of personnel psychologists to make wide use of these equations are presented and shown to be faulty. Decision theoretic equations are used to estimate the impact on productivity of a valid test if used to select new computer programers for one year in (a) The Federal Government and (b) the national economy. The test analyzed is the Programer Aptitude Test (PAT), which previous validity generalization research (Rosenberg, Schmidt, and Hunter, Note 1) has shown to have substantial and generalizable validity. A newly developed technique is used to estimate SD_y, the standard deviation of the dollar value of employee job performance, the item of required information that has been most difficult and expensive to estimate in the past. Results are presented separately for the Federal Government and U.S. economy. For both, results are presented for different selection ratios and for different assumed values for the validity of previously used selection procedures. The impact of PAT on programer productivity was shown to be substantial for all combinations of assumptions. The results support the conclusion that hundreds of millions of dollars in increased productivity could be realized by increasing the validity of selection decisions in this occupation. Likely similarities between computer programers and other occupations are also discussed.

The Impact of Valid Selection Procedures on Workforce Productivity

Questions concerning the economic and productivity implications of valid selection procedures have come increasingly to the fore in industrial-organizational psychology. The recent Annual Review of Psychology chapter by Dunnette and Borman (1979) includes—for the first time—a separate section on the utility and productivity implications of selection methods. This development is due at least in part to the emphasis placed on the practical utility of selection procedures in recent years in some of the litigation involving selection tests. Hunter and Schmidt (1979) have contended, on the basis of a review of the empirical literature on the economic utility of selection procedures, that personnel psychologists have typically failed to appreciate the magnitude of productivity gains that result from use of valid selection procedures. The major purpose of this study is to illustrate the productivity (economic utility) implications of a valid selection procedure in the occupation of computer programer in the Federal Government and in the economy as a whole.

History and Development of Selection Utility Models

The evaluation of benefit obtained from selection devices has been a problem of continuing interest in industrial psychology. Most attempts to evaluate benefit have focused on the validity coefficient, and at least five approaches to the interpretation of the validity coefficient have been advanced over the years. The oldest of these is the Index of Forecasting Efficiency, symbolized E. $E = 1 - \sqrt{1 - r_{xy}^2}$, where r_{xy} is the validity coefficient. This index compares the standard error of job performance scores predicted by means of the test (the standard error of estimate) to the standard error that results when there is no valid information about applicants and one predicts the mean level of performance for everyone (the standard deviation of job performance). The index of forecasting efficiency was heavily emphasized in early texts (Kelley, 1923; Hull, 1928) as the appropriate means for evaluating the value of a selection procedure. This index describes a test correlating .50 with job performance as predicting only 13% better than chance, a very unrealistic and pessimistic interpretation of the economic test's value.

The index of forecasting efficiency was succeeded by the coefficient of determination, which became popular during the 1930's and 1940's. The coefficient of determination is simply the square of the validity coefficient or r_{xy}^2. This coefficient was refered to as "the proportion of variance in the job performance measure accounted" for by the test. The coefficient of determination describes a test of validity of .50 as "accounting for" 25% of the variance of job performance. Although r_{xy}^2 is still occasionally referred to by selection psychologists--and has surfaced in litigation on personnel tests--the "amount of variance accounted for" has no direct relationship to productivity gains resulting from use of selection device.

Both E and r_{xy}^2 lead to the conclusion that only tests with relatively high correlation with job performance will have significant practical value. Neither of these interpretations recognizes that the value of a test varies as a function of the parameters of the situation in which it is used. They are general interpretations of the correlation coefficient and have been shown to be inappropriate for interpreting the validity coefficient in selection (Brogden, 1946; Cronbach and Gleser, 1965, p. 31; Curtis and Alf, 1969).

The well-known interpretation developed by Taylor and Russell (1939) goes beyond the validity coefficient itself and takes into account two properties of the selection problem--the selection ratio (the proportion of applicants hired) and the base rate (the percentage of applicants who would be "successful" without use of the test). This model yields a much more realistic interpretation of the value of selection devices. The Taylor-Russell model indicates that even a test with a modest validity can substantially increase the percentage who are successful among those selected when the selection ratio is low. For example, when the base rate is .50 percent and the selection ratio is .10, a test with validity of only .25 will increase the percentage among the selectees who are successful from 50 percent to 67 percent, a gain of 17 additional successful employees per 100 hired. Although an improvement, the Taylor-Russell approach to determining selection utility does have disadvantages. Foremost among them is the need for a dichotomous criterion. Current employees and new hires must be sorted into an unrealistic two point distribution of job performance: "successful" and "unsuccessful" (or "satisfactory" and "unsatisfactory"). The decision as to where to draw the line to create the dichtomy is arbitrary. But more important than this is the fact that information on levels of performance within each group is lost (Cronbach & Gleser, 1965, 123-124, 138). All those within the "successful" group, for example, are implicitly assumed equal in value, whether they perform in an outstanding manner or barely exceed the cut-off. This fact makes it difficult to express utility in units that are comparable across situations.

The next major advance was left to Brogden (1949), who used the principles of linear regression to demonstrate how the selection ratio (SR) and the standard deviation of job performance in dollars (SD_y) affect the economic utility of a selection device. Despite the fact that Brogden's derivations are a landmark in the development of selection utility models, they are very straightforward and simple to understand.

Let r_{xy} = the correlation between the test (x) and job performance measured in dollar value. The basic linear model is:

$$Y = \beta Z_x + \mu_y + e$$

Where:

Y = job performance measured in dollar value;

β = the linear regression weight on test scores for predicting job performance;

Z_x = test performance in standard score form in the applicant group;

μ_y = mean job performance (in dollars) of randomly selected employees; and

e = error of prediction.

This equation applies to the job performance of an individual. The equation which gives the average job performance for the selected (s) groups (or for any other subgroup) is:

$$E(Y_s) = E(\beta Z_{x_s}) + E(\mu_y) + E(e)$$

Since E(e) = 0, and β and μ are constants, this becomes:

725

$$\bar{Y}_s = \beta\bar{Z}_{x_s} + \mu_y$$

This equation can be further simplified by noting that $\beta = r_{xy}\ (SD_y/SD_x)$ where SD_y is the standard deviation of job performance measured in dollar value among randomly selected employees. Since $SD_x = 1.00$, $\beta = r_{xy}SD_y$. We thus obtain:

$$\bar{Y}_s = r_{xy}SD_y\bar{Z}_{x_s} + \mu_y$$

This equation gives the absolute dollar value of average job performance in the selected group. What is needed is an equation which gives the increase in dollar value of average performance that results from using the test. Note that if the test were not used, $\bar{Y}_s$ would be μ_y. That is, mean performance in the selected group is the same as mean performance in a group selected randomly from the applicant pool. Thus the increase due to use of a valid test is $r_{xy}\ SD_y\ \bar{Z}_{x_s}$. The equation we want is produced by transposing μ_y to give:

$$Y_s - \mu_y = r_{xy}SD_y\bar{Z}_{x_s}$$

The value on the right in the above equation is the difference between mean productivity in the group selected using the test and mean productivity in a group selected without using the test, that is, a group selected randomly. The above equation thus gives mean gain in productivity per selectee resulting from use of the test, i.e.,

$$\Delta\bar{U}/\text{selectee} = r_{xy}SD_y\bar{Z}_{x_s} \qquad (1)$$

where U is utility and ΔU is marginal utility.

Equation (1) states that the average productivity gain in dollars per person hired is the product of the validity coefficient, the average standard score on the test of those hired, and the SD of job performance in dollars. The value $r_{xy}\bar{Z}_{x_s}$ is the mean standard score on the dollar criterion of those selected, Z_y. Thus utility per selectee is the mean Z-score on the criterion of those selected times the standard deviation of the criterion in dollars. The only assumption that Equation (1) makes is that the relation between the test and job performance is linear. If we further assume that the test scores are normally distributed, the mean test score of those selected is ϕ/p, where:

p = the selection ratio, and

ϕ = the ordinate in $N(0,1)$ at the point of cut corresponding to p.

Thus equation [1] can be written:

$$\Delta U/\text{selectee} - r_{xy}\ \phi/p\ SD_y \qquad (2)$$

The above equations illustrate the critical role of SD_y and suggests the possibility of situations in which tests of low validity have higher utility than tests of high validity. For example:

	r_{xy}	$\overline{Z}_x$	SD_y	ΔU/selectee
Mid-level job (e.g., systems analyst)	.20	1.00	25,000	$5,000
Lower level job (e.g., janitor)	.60	1.00	2,000	1,200

The total utility of the test depends on the number of persons hired. The total utility (total productivity) gain resulting from use of the test is simply the mean gain per selectee times the number of people selected, N_s. That is, the total productivity gain is:

$$\Delta U = N_s\ r_{xy}\ SD_y\ \overline{Z}_{x_s}.$$

In this example, the average, marginal utilities are $5000 and $1200. If 10 people were hired the actual utilities would be $50,000 and $12,000 respectively. If 1000 people were to be hired, then the utilities would be

\$500,000 and \$120,000 respectively. Obviously the total dollar value of tests is greater for large employers than for samll employers. However, this fact can be misleading: on a percentage basis it is average gain in utility that counts; and that's what counts to each individual employer.

Equations (1) and (2) clearly illustrate the basis for Brogden's (1946) conclusion that the validity coefficient itself is a direct index of selective efficiency. Brogden (1946) showed that, given only the assumption of linearity, the validity coefficient is the proportion of maximum utility attained, where maximum utility is the productivity gain that would result from a perfectly valid test. A test with a validity of .50, for example, can be expected to produce 50% of the gain that would result from a perfect (validity = 1.00) selection device used in the same setting and at the same selection ratio. A glance at Equation (1) or Equation (2) verifies this verbal statement. Since the validity coefficient enters the equation as a multiplicative factor, increasing or decreasing the validity by any factor will increase or decrease the utility by the same factor. For example, if we increase validity by a factor of two by raising it from .20 to .40, equation (2) shows that utility doubles. If we decrease validity by a factor of one-half by lowering it from 1.00 to .50, utility is cut in half. Equations (1) to (2) also illustrate the fact that there are limitations on the utility of even a perfectly valid selection device. If the selection ratio is very high, the term ϕ/p (or $\bar{z}_{x_s}$) approaches zero and even a perfect test has little value. If the selection ratio is 1.00, the perfect test has no value at all. Likewise, as SD_y decreases, the utility of even a perfect test decreases. In a hypothetical world in which SD_y were zero, even a perfect test would have no value.

Brogden (1946) further showed that the validity coefficient could be expressed as the following ratio:

$$r_{xy} = \frac{\overline{Z}_{y(x)} - \overline{Z}_{y(r)}}{\overline{Z}_{y(y)} - \overline{Z}_{y(r)}}$$

where:

$\overline{Z}_{y(x)}$ = the mean job performance (y) standard score for those selected using the test (x).

$\overline{Z}_{y(y)}$ = the mean job performance standard score resulting if selection were on the criterion itself, at the same selection ratio.

$\overline{Z}_{y(r)}$ = the mean job performance standard score resulting if selection decisions were made randomly (from among the otherwise screened pool of applicants).

r_{xy} = the validity coefficient.

Since $\overline{Z}_{y(r)} = 0$ by definition, the above formula reduces to $\overline{Z}_{y(x)}/\overline{Z}_{y(y)}$. This formulation has implications for the development of new methods of estimating selection procedure validity. If reasonably accurate estimates of both $\overline{Z}_{y(x)}$ and $\overline{Z}_{y(y)}$ can be obtained, validity can be estimated without conducting a traditional validity study. Further, estimates produced by a procedure of this kind would be unaffected by range restriction and criterion unreliability.

In Equations (1) and (2), the values for r_{xy} and SD_y should be those which would hold if applicants were hired randomly with respect to test scores. That is, they should be values applicable to the applicant population, the group in which the selection procedure is actually used. Values of r_{xy} and SD_y computed on incumbents will typically be underestimates because of reduced variance among incumbents on both test and job performance measures. Values of r_{xy} computed on incumbents can be corrected for range restriction to produce estimates of the value in the applicant pool (Thorndike, 1949, 169-176). The applicant pool is made up of all who have survived screening on any prior selection hurdles than might be employed, e.g., minimum educational requirements, physical examinations, etc.

The correlation between the test and a well developed measure of job performance (y') provides a good estimate of r_{xy}, the correlation of the test with job performance measured in dollars (productivity). It is a safe assumption that job performance and the value of that performance are at least monotonically related. It is inconceivable that lower performance could have greater dollar value than higher performance. Ordinarily, the relation between y' and y will be not only monotonic but also linear. If there are departures from linearity, the departures will typically be produced by leniency in job performance ratings which lead to ceiling effects in the measuring instrument. The net effect of such ceiling effects is to make the test's correlation with the measure of job performance smaller than its correlation with actual performance, that is, smaller than its true value, making $r_{xy'}$ an underestimate of r_{xy}. An alternative statement of this effect is that ceiling effects due to leniency produce an artificial nonlinear relation between job performance ratings and the actual dollar value of performance. A nonlinear relation of this form would lead to an underestimation of selection utility because the performance measure underestimates the relative value of very high performers. Values of $r_{xy'}$ should also be corrected for attenuation due to errors of measurement in the criterion. Random error in the observed measure of job performance causes the test's correlation with that measure to be lower than its correlation with actual job performance. Since it is the correlation with actual performance that determines test utility, it is the attenuation-corrected estimate that is needed in the utility formulas. This estimate is simply $r_{xy}/\sqrt{r_{y'y'}}$ where $r_{y'y'}$ is the reliability of the performance measure. (See Schmidt, Hunter, & Urry, 1976, for further discussion of these points.)

The next major advance in this area came in the form of the monumental work by Cronbach and Gleser, Psychological Tests and Personnel Decisions. First published in 1954, this work was republished in 1965 in augmented form. The book consists of detailed and sophisticated application of decision theory principles not only to the single-stage fixed-job selection decisions which we have thus far discussed, but also to placement and classification decisions and sequential selection strategies. In these latter areas, many of their derivations were indeed new to the field of personnel testing. Their formulas for utility in the traditional selection setting, however, turn out upon examination to be identical to those of Brogden (1949), except for the fact that they formally incorporate cost of testing ("information gathering") into the equations.

Brogden, it will be recalled, approached the problem from the point of view of mean gain in utility per selectee. Cronbach and Gleser (1965, chapter 4) derived their initial equation in terms of mean gain per applicant. Their initial formula was (ignoring cost of testing for the moment):

$$\Delta U/\text{applicant} = r_{xy}\, SD_y \phi$$

All terms are as defined earlier. Mulitplying by the number of applicants, N, yields total or overall gain in utility. The Brogden formula for overall utility is:

$$\Delta U = N_S\ \Delta\overline{U}/\text{selectee} = N_S\ r_{xy}\ SD_y\ \phi/p \qquad (3)$$

Ns, it will be recalled, is the number selected. If we note that $p = N_S/N$, i.e., the ratio of selectees to applicants, we find that Brogden's equation immediately reduces to the Cronbach-Gleser (1965) equation for total utility:

$$\Delta U = N r_{xy} SD_y\ \phi$$

Role of the Cost of Testing

The previous section ignored the cost of testing, which is quite reasonable in most testing situations. For example, in a typical job situation, the applicant pool consists of people who walk through the door and ask for a job (i.e., there are no recruiting costs). Hiring is then done on the basis of an application blank and a test which are administered by a trained clerical worker at a cost of 10 dollars or so. If the selection ratio is 10%, then the cost of testing per person hired is 10 dollars for each person hired and 90 dollars for the nine persons rejected in finding the person hired, or 100 dollars altogether. This is negligible in relation to the usual magnitude of utility gains. Furthermore, this 100 dollars is a one time cost whereas utility gains continue to accumulate over as many years as the person hired stays with the organization. When cost of testing is included, Equation (2) becomes:

$$\Delta\overline{U}/\text{selectee} = r_{xy}\ SD_y\ \phi/p - C/p \qquad (4)$$

where C is the cost of testing one applicant.

Although cost of testing typically has only a trivial impact on selection utility, it is possible to conjure up hypothetical situations in which cost plays a critical role. For example, suppose an employee were recruiting one individual for a sales position that would last only one year. Suppose further that the employers decide to base their selection on the results of an assessment center which costs $1000 per assessee and has a true validity of .40. If the yearly value of SD_y for this job is $10,000, and 10 candidates are assessed, the expected gain in productivity is .4($10,000)(1.758) or $7034. However, the cost of the assessment center is 10(1000) = $10,000, which is $2966 greater than the expected productivity gain. That is, under these conditions it would cost more to test 10 persons than would be gained in improved performance. If the employer tested only five candidates, then the expected gain in performance would be 5607 dollars while the cost of testing would be $5000 for an expected gain of 607 dollars. In this situation, the optimal number to test is three persons. The best person of three would have an expected gain in performance of $4469 with a cost of testing of 3000 dollars, for an expected utility of 1469 dollars.

Relation Between SR and Utility

In most situations, the number to be hired is fixed by organizational needs. If the applicant pool is also fixed, the question of which SR would yield maximum utility becomes academic. The SR is determined by circumstances and is not under the control of the employer. However, employers can often exert some control over the size of the applicant pool by increasing or decreasing recruiting efforts. If this is the case, the question is then how many applicants the employer should test to obtain the needed number of new employees in order to maximize productivity gains from selection. This question can be answered using a formula given by Cronbach and Gleser (1965, p. 309):

$$\phi - pZ_x = C/r_{xy} SD_y$$

where Z_x is the cutting score on the test in Z score form. This equation must be solved by iteration. Only one value of the SR (i.e., p) will satisfy this equation and p will always be less than or equal to .50. The value computed for the optimal SR indicates the number that should be tested in relation to the number to be selected. For example, if the number to be selected is 100 and Equation (3) indicates that the optimal SR is .05, the employer will maximize selection utility by recruiting and testing 2000 candidates (100/.05 = 2000). The cost of recruiting additional applicants beyond those available without recruitment efforts must be incorporated into the cost of testing term, C. C then becomes the average cost of recruiting and testing one applicant. The lower the cost of testing and recruiting, the larger the number of applicants it is profitable to test in selecting a given number of new employees. Since the cost of testing is typically quite low relative to productivity gains from selection, the number tested should typically be large relative to the number selected.

In situations in which the applicant pool is constant, statements about optimal SR's typically do not have practical value, since the SR is not under the control of the employer. Given a fixed applicant pool, $\Delta\overline{U}$/selectee increases as SR ratio decreases if cost of testing is not considered. Brogden (1949) showed that, when cost of testing is taken into account and when this cost is unusually high, $\Delta\overline{U}$/selectee will be less at very low SR's than at somewhat high SR's. If cost of testing per applicant is very high, cost of testing per selectee can become greater at extremely low SR's than $\Delta\overline{U}$/selectee, producing a loss rather than a gain in utility. In practice, however, the combination of extremely high testing costs and extremely low SR's that could lead to negative utilities occurs rarely, if ever. When the applicant pool is fixed, the SR that is optimal for $\Delta\overline{U}$/selectee is not necessarily the optimal SR for total gain in utility. Cronbach and Gleser showed that total utility is always greatest when the SR falls at .50. As SR decreases from .50, $\Delta\overline{U}$/selectee increases until it reaches its maximum, the location of which depends on the cost of testing. But as $\Delta\overline{U}$/selectee increases, the number of selectees, N_S, is decreasing, and the product $N_S\Delta\ \overline{U}$/selectee or total utility is also decreasing. In a fixed applicant pool, total gain is always greatest when 50 percent are selected and 50 percent are rejected (Cronbach and Gleser, 1965, pp. 38-40).

Reasons for Failure to Employ Selection Utility Models

Despite the availability since 1949 of the utility equations discussed above, applied differential psychologists have been notably slow in carrying out decision-theoretic utility analyses of selection procedures. In our judgement, the sparcity of work in this area is primarily traceable to three facts. First, many psychologists believe that the utility equations presented above are of [illegible] value unless the data exactly fit the linear homoscedastic model and all marginal distributions are normal. Many reject the model in the belief that their data do not perfectly meet the assumptions.

Second, psychologists once believed that validity is situationally specific, that there are subtle differences in the performance requirements of jobs from situation to situation that produce (nontrivial) differences in test validities. If this were true, then the results of a utility analysis conducted in a given setting could not be generalized to apparently identical test-job combinations in new settings. Combined with the belief that utility analyses must include costly cost accounting applications, it is easy to see why belief in situational specificity of test validities would lead to reluctance to carry out utility analyses.

Third, it has been extremely difficult in most cases to obtain all the information called for by the equations. The SR and cost of testing can be determined reasonably accurately and at relatively little expense. The item of information that has been most difficult to obtain is the needed estimate of SD_y (Cronbach & Gleser, 1965, p. 121). It has generally been assumed that SD_y can be estimated only by the use of costly and complicated cost accounting methods. These procedures involve first costing out the dollar value of the job behaviors of each employee (Brogden & Taylor, 1950) and then computing the standard deviation of these values. In an earlier review (Hunter & Schmidt, 1978), we were able to locate only two studies in which cost accounting procedures were used to estimate SD_y. In this study, we will present an alternative to cost accounting estimates of SD_y.

Are the Statistical Assumptions Met?

The linear homoscedastic model includes three assumptions:

1. Linearity.
2. Equality of variances of conditional distributions.
3. Normality of conditional distributions.

736

As we have shown above, the basic selection utility equation [Equation (1)] depends only on linearity. Equation (2) does assume normality of the test score distribution. However, Brogden (1949) and Cronbach and Gleser (1965) introduced this assumption essentially for derivational convenience: it provides an exact relation between the SR and $\overline{Z}_{x_s}$. One need not use the normality-based relation $\phi/p = \overline{Z}_{x_s}$ to compute $\overline{Z}_{x_s}$. The value of $\overline{Z}_{x_s}$ can be computed directly. Thus in the final analysis, linearity is the only required assumption.

To what extent does data in differential psychology fit the linear homoscedastic model? To answer this question, we must of necessity examine sample rather than population data. However, it is only conditions in populations that are of interest; sample data is of interest only as a means of inferring the state of nature in populations. Obviously, the larger the sample used, the more clearly the situation in the sample will reflect that in the population, given that the sample is random. A number of researchers have addressed themselves to this problem.

Sevier (1957), using N's from 105 to 250, tested the assumptions of linearity, normality of conditional criterion distributions, and equality of conditional variances. The data were from an education study, with cumulative grade point average being the criterion and high school class rank and various test scores being the predictors. Out of 24 tests of the linearity assumption, only one showed a departure significant at the .05 level. Out of 8 samples tested for equality of conditional variances, only one showed a departure significant at the .05 level. However, 25 of the 60 tests for normality of the conditional criterion distributions were significant at the .05 level. Violation of this assumption throws interpretations of conditional standard deviations based on normal curve tables into some doubt. However, this statistic is typically not used in practical prediction situations, such as selection or placement. Sevier's study indicates that the assumptions of linearity and equality of conditional variances may be generally tenable.

Ghiselli and Kahneman (1962) examined 60 aptitude varibles on one sample of 200 cases and reported that fully 40 percent of the variables departed significantly from the linear homoscedastic model. Ninety percent of these departures were reported to have held up on cross-validation. Tupes (1964) re-analyzed the Ghiselli and Kahneman data and found that only 20 percent of the relationships departed from the linear honoscedastic model at the .05 level. He also found that three of the "significant" departures from linearity were probably due to typographical or clerical errors in the data. Later Ghiselli (1964) accepted and agreed with Tupes' re-analysis of his data. Tupes' findings must be interpreted in light of the fact that the frequency of departure from the linear homoscedastic model expected at the .05 level is in fact much greater than 5%. Tupes carried out two statistical tests on each test-criterion relation: one for linearity and one for equality of conditional variances. Thus the expected proportion of data samples in which at least one test is significant is not .05 but rather a little over .09. If three statistical tests are run at the .05 level—one for linearity, one for normality of conditional distributions, and one for homogeneity of conditional distributions, the expected proportion of data samples in which at least one of these tests is significant is approximately .14 when relations in the parent populations are perfectly linear and homoscedastic.

Tiffin and Vincent (1960) found no significant departures from the bivariate normal model in 15 independent samples of test-criterion data, ranging in size from 14 to 157. In each set of data, a chi square test was used to compare the percent of employees in the "successful" job performance category in each fifth of the test score distribution to the percentages predicted from the normal bivariate surface (which incorporates the linear homoscedastic model) corresponding to the computed validity coefficient. Surgent (1947) performed a similar analysis on similar data and reported the same findings.

Hawk (1970) reported a major study researching for departures from linearity. The data were drawn from 367 studies conducted on the General Aptitude Test Battery (GATB), used by the U.S. Department of Labor, between 1950 and 1966. A total of 3303 relations, based on 23,428 individuals, between the nine subtests of the GATB and measures of job performance (typically supervisory ratings) were examined. The frequency of departures from linearity significant at the .05 level was .054. Using the .01 level, the frequency was .012. Frequencies closer to the chance level can hardly be imagined.

Brogden, during his years as technical director of what is now the Army Research Institute for the Behavior and Social Sciences, spent a considerable amount of time and effort attempting to identify nonlinear test-criterion relationships in large samples of military selection data.

Although quadratic and other higher order nonlinear equations sometimes provided impressive fits to the data in the initial sample, not one of the equations cross-validated successfully in a new sample from the same population. In cross-validation samples, the nonlinear functions were never superior to simple linear functions (Brogden, Note 1).

These findings, taken in toto, indicate that the linear homoscedastic model generally fits the data in this area quite well. The linearity assumption, the only truly critical assumption, is particularly well supported.

We turn now to the question of normality of marginal distributions. In certain forms (see Equation 2), the Brogden-Cronbach utility formulas assume, in addition to linearity, a normal distribution for predictor (test) scores. The Taylor-Russell tables, based on the assumption of a normal bivariate surface, assume normality of total test score distribution also. One obviously relevant question is whether or not violations of this assumption seriously distort utility estimates. Van Naersson (in Cronbach & Gleser, 1965) found that they do not. He derived a set of utility equations parallel to the Brogden-Cronbach equations except that they were based on the assumption of a rectangular distribution of test scores. He found that when applied to the same set of empirical data, the two kinds of equation produced very similar utility estimates (p. 288). Cronbach and Gleser (1965, p. 160) point out that this finding "makes it possible to generalize over the considerable variety of distributions intermediate between normal and rectangular." Results from the Schmidt and Hoffman (1973) study suggest the same conclusion. In their data neither the predictor nor the criterion scores appeared to be normally distributed. Yet the utility estimates produced by the Taylor-Russell tables were only off marginally: 4.09 percent at SR = .30 and 11.29 percent at SR = .50.

Thus it appears that an obsessive concern with statistical assumptions is not justified. This is especially true in light of the fact that for most purposes, there is no need for utility estimates to be accurate down to the last dollar. Approximations are usually quite adequate for the kinds of decisions that these estimates are used to make (Van Naersson, 1963, p. 282; cf. also Cronbach & Gleser, 1965, 139). Alternatives to use of the utility equations will typically be procedures which produce larger errors, or even worse, no utility analyses at all. Faced with these alternatives errors in the 5-10 percent range appear negligible. Further, if overestimation of utility is considered more serious than underestimation, one can always employ conservative estimates of equation parameters (e.g., r_{xy}, SD_y) to virtually guarantee against overestimation of utilities.

Are Test Validities Situationally Specific?

The third reason we postulated for the failure of personnel psychologists to exploit the Brogden-Cronbach utility models was belief in the doctrine of situational specificity of validity coefficients. This belief precludes generalization of validities from one setting to another, making criterion-related validity studies—and utility analyses—necessary in each situation. The empirical basis for the principle of situational specificity has been the fact that considerable variability in observed validity coefficients is typically apparent from study to study even when jobs and tests appear to be similar or essentially identical (Ghiselli, 1966). However, there are a priori grounds for postulating that this variance is due to statistical, measurement, and other artifacts unrelated to the underlying relation between test and job performance. There are at least seven such sources of artifactual variance:

1. Differences between studies in criterion reliability.
2. Differences between studies in test reliability.
3. Differences between studies in range restriction.
4. Sampling error (i.e., variance due to $N < \infty$).
5. Differences between studies in amount and kind of criterion contamination and deficiency (Brogden & Taylor, 1950).
6. Computational and typographical errors (Wolins, 1962).
7. Slight differences in factor structure between tests of a given type (e.g., arithmetic reasoning tests).

In a purely analytical substudy, Schmidt et al. (in press) showed that the first four sources alone are capable, under specified and realistic circumstances, of producing as much variation in validities as is typically observed from study to study. They then turned to analyses of empirical data. Using 14 distributions of validity coefficients from the published and unpublished literature for various tests in the occupations of clerical worker and first-line supervisor, they found that artifactual variance sources (1) through (4) accounted for an average of 62 percent of the variance in validity coefficients, with a range from 43 percent to 87 percent. Thus there was little remaining variance in which situational moderators could operate. In an earlier study (Schmidt & Hunter, (1977), it was found that sources (1), (3) and (4) alone accounted for an average of about 50 percent of the observed variance in distributions of validity coefficients presented by Ghiselli (1966, p. 29). If one could correct for all seven sources of error variance, one would, in all likelihood, consistently find that the remaining variance was zero or near zero. That is, it is likely that the small amounts of remaining variance in the studies cited here are due to the sources of artifactual variance not corrected for. Thus there is now strong evidence that the observed variation in validities from study to study for similar test-job combinations is artifactual in nature. These findings cast considerable doubt on the situational specificity hypothesis.

Rejection of the situational specificity doctrine obviously opens the way to validity generalization. However, validity generalization is possible in many cases even if the situational specificity hypothesis cannot be definitively rejected. After correcting the mean and variance of the validity distribution for sampling error, for attenuation due to criterion unreliability, and for range restriction (based on average values of both), one may find that a large percentage, say 90 percent, of all values in the distribution lie above the minimum useful level of validity. In such a case, one can conclude with 90% confidence that true validity is at or above this minimum level in a new situation involving the same test-type and job without carrying out a validation study of any kind. Only a job analysis is necessary, in order to ensure that the job at hand is a member of the class of jobs on which the validity distribution was derived. In Schmidt and Hunter (1977), two of the four validity distributions fell into this category, even though only three sources of artifactual variance could be corrected for. In the later study (Schmidt et al., in press) in which it was possible to correct for four sources of error variance, 12 of the 14 corrected distributions had 90 percent or more of validities above levels that would typically be indicative of significant practical utility (cf. Hunter & Schmidt, 1979).

These methods and findings indicate that in the future validity generalization will be possible for a wide variety of test-job combinations. Such a development will do much to encourage the application of decision-theoretic utility estimation tools.

Difficulties in Estimating SD_y

The third major reason for neglect of the powerful Brogden-Cronbach utility model was the difficulty of estimating SD_y. As noted above, the generally recommended procedure for estimating SD_y is by use of cost accounting procedures. Such procedures are supposed to be used to estimate the dollar value of performance of a number of individuals (cf. Brogden & Taylor, 1950a), and the SD of these values is then computed. Roche's (1961) dissertation illustrates well the tremendous time and effort such an endeavor entails. This study (summarized in Cronbach and Gleser, 1965, pp. 256-266) was carried out on radial drill operators in a large midwestern plant of a heavy equipment manufacturer. A cost accounting procedure called "standard costing" was used to determine the contribution of each employee to the profits of the company. The procedure was extremely detailed and complex, involving cost estimates for each piece of material machined, direct and indirect labor costs, overhead, perishable tool usage, etc. There was also a "burden adjustment" for below standard performance. But despite the complexity and apparent objectivity, Roche is compelled to admit that "many estimates and arbitrary allocations entered into the cost accounting" (p. 263, in Cronbach & Gleser, 1965). Cronbach, in commenting on the Roche study after having discussed it with Roche, states that some of the cost accounting procedures used are unclear or questionable (Cronbach & Gleser, 1965, pp. 266-267) and that the accountants perhaps did not fully understand the utility estimation problem. Thus even given great effort and expense, cost accounting procedures may nevertheless lead to a questionable final product.

Recently we have developed a procedure for obtaining rational estimates of SD_y. This method was used in a pilot study by 62 experienced supervisors of budget analysts to estimate SD_y for that occupation. Supervisors were used as judges because they have the best opportunities to observe actual performance and output differences between employees on a day-to-day basis. The method is based on the following reasoning: if job performance in dollar terms is normally distributed, then the difference between the value to the organization of the products and services produced by the average employee and those produced by an employee at the 85th percentile in performance is equal to SD_y. Budget Analyst supervisors were asked to estimate both these values; the final estimate was the average difference across the 62 supervisors. The estimation task presented to the supervisors may appear difficult at first glance, but only one out of 62 supervisors objected and stated that he did not think he could make meaningful estimates. Use of a carefully developed questionnaire to obtain the estimates apparently aided significantly; a similar questionnaire was used in the present study and is described later. The final estimate of SD_y for the budget analyst occupation was 11,327 per year (standard error of the mean = \$1,120). This estimate is based on incumbents rather than applicants and must therefore be considered to be an underestimate. As noted earlier, it is generally not critical that estimates of utility be accurate down to the last dollar. Utility estimates are typically used to make decisions about selection procedures, and for this purpose only errors large enough to lead to incorrect decisions are of any consequence. Such errors may be very infrequent. Further, they may be as frequent—or more frequent—when cost accounting procedures are used. As we noted above, Roche (1961) found that, even in the case of the simple and structured job he studied, the cost accountants were frequently forced to rely on subjective estimates and arbitrary allocations. This is generally true in cost accounting and may become a more severe problem as one moves up the occupational hierarchy. What objective cost accounting techniques, for example, can be used to assess the dollar value of an executive's impact on subordinate morale? It is the jobs with the largest SD_y values, i.e., the jobs for which $\Delta\overline{U}$/selectee is potentially greatest, that are handled least well by cost accounting methods. Rational estimates—to one degree or another—are virtually unavoidable at the higher job levels.

Our procedure has at least two advantages in this respect. First, the mental standard to be used by the supervisor-judges is the estimated cost to the organization of having an outside consulting firm provide the same products and/or services. In many occupations, this is a relatively concrete standard. Second, the idiosyncratic tendencies, biases, and random errors of individual experts can be controlled for by averaging across a large number of judges. In our initial study, the final estimate of SD_y was the average across 62 supervisors. Unless this is an upward or downward bias in the group as a whole, such an average should be fairly accurate. In our example, the standard error of the mean was 1,120. This means that the interval $9,480 to $13,175 should contain 90 percent of such estimates. (One truly bent on being conservative could employ the lower bound of this interval in his or her calculations.)

Methods similar to the one described here have been used successfully by the Decision Analysis Group of the Stanford Research Institute (Howard, Note 2) to scale otherwise unmeasurable but critical variables. Resulting measures have been used in the application of decision-theoretic principles to high-level policy decision-making in such areas as nuclear power plant construction, corporate risk policies, investment and expansion programs, and hurricane seeding (Howard, 1966; Howard & Matheson, 1972, Raffia, 1968; Matheson, Note 3). All indications are that the response to the work of this group has been quite positive; these methods have been judged by high level decision-makers to contribute valuably to improvement of socially and economically important decisions.

In most cases, the alternatives to use of a procedure like ours to estimate SD_y are unpalatable. The first alternative is to abandon the idea of a utility analysis. This course of action will typically lead to a gross (implicit) underestimate of the economic value of valid selection procedures. This follows if one accepts our contention (Hunter & Schmidt, 1979) that the empirical studies that are available indicate much higher dollar values than psychologists have expected. The second alternative in most situations is use of a less systematized, and probably less accurate, procedure for estimating SD_y. Both these alternatives can be expected to lead to more erroneous decisions about selection procedures.

The procedure for estimating SD_y described here assumes that dollar outcomes are normally distributed. One purpose of the present study is to evaluate that assumption.

The present study has three purposes: (1) to illustrate the magnitude of the productivity implications of a valid selection procedure, (2) to demonstrate the application of decision-theoretic utility equations, and (3) to test the assumption that the dollar value of employee productivity is normally distributed.

Procedure

The major reason for our choice of the job of computer programer was that a previous study (Rosenberg, Schmidt, & Hunter, Note 4) had provided remarkably accurate validity estimates for this job. Applying the Schmidt-Hunter (1977) validity generalization model to all available validity data for the Programer Aptitude Test (PAT; Hughes & McNamara, Note 5; McNamara & Hughes, 1961), this study found that the percent of variance in validity coefficients accounted for in the case of job proficiency criteria for the PAT total score was 94 percent. This finding effectively refutes the situational specificity hypothesis. The estimated true validity was .76.

Thus the evidence is strong that the (multivariate) total PAT score validity is quite high for predicting performance of computer programers and that this validity is essentially constant across situations (e.g., different organizations; Rosenberg, et al. Note 4). Since it is total score that is typically used in selecting programers, this study concerns itself only with total score validity. Because the PAT is no longer available commercially, testing costs had to be estimated. In this study, we assumed a testing cost of $10 per examinee.

Estimates of SD_y were provided by experienced supervisors of computer programers in 10 Federal agencies. These supervisors were selected by their own supervisors after consultation with the first author. Participation was voluntary. Of 147 questionnaires distributed, 105 were returned (all in usable form), for a return rate of 71.4%. In order to test the hypothesis that dollar outcomes are normally distributed, the supervisors were asked to estimate values for the 15th percentile ("low performing programers"), as well as the 50th percentile ("average programers"), and the 85th percentile ("superior programers"). The resulting data thus provides two estimates of SD_y. If the distribution is approximately normal, these two estimates will not differ substantially in value.

The questionnaire used to elicit supervisor estimates is shown (in relevant part) in Table 1. The wording of this questionnaire was carefully developed and pretested on a small sample of programer supervisors and personnel psychologists. None of the programer supervisors who returned questionnaires in the study reported any difficulty understanding the questionnaire or in making the estimates.

This study focuses on selection of computer programers at the GS-5 through 9 levels. GS level 5 is the lowest level in this occupational series. Beyond GS-9, it is unlikely that an aptitude test like the PAT would be used in selection. Applicants for higher level programer positions are expected (and required) to have considerable developed expertise in programing, and are selected on the basis of achievement and experience, rather than directly on aptitude. The vast majority of programers hired at the GS-9 level are promoted to GS-11 after one year. Similarly all but a minority hired at the GS-5 level advance to GS-7 in one year and to GS-9 the following year. Therefore the SD_y estimates were obtained for the GS-9-11 levels, as can be seen in Table 1. Statistical information obtained from the Bureau of Personnel Management Information Systems of the Civil Service Commission indicated that the number of programer incumbents in the Federal Government at the relevant levels (GS-5 through 9) was 4,404 (as of October 31, 1976, the latest date for which figures were available). The total number of computer programers at all grade levels was 18,498. For 1975-1976, 61.3 percent of all new hires were at the GS-5-9 levels. The number of new hires government-wide in this occupation at these levels was 655 for 565 for calendar years 1975 and 1976, respectively, for an average yearly selection rate of 618. The average tenure of the GS-5-9 computer programers was determined to be 9.69 years.

Data from the 1970 U.S. Census showed that there were 166,556 computer programers in the U.S. in that year. Because the growth rate has been rapid in this occupation recently, this figure undoubtedly underestimates the current number of programers. However, it is the most most recent estimate available. In any event, the effect of underestimation on the utility results is a conservative one. It was not possible to determine the number of computer programers that are hired yearly in the U.S. economy. For purposes of this study, it was assumed that the turnover rate was 10 percent in this occupation and that therefore .10 (166,556) or 16,655 were hired to replace those who had quit, retired or died. Extrapolating from the Federal to the private sector workforce, it was assumed that 61.3 percent of these new hires were at occupational levels for which the PAT would be appropriate. Thus it was assumed that .613 (16,655) or 10,210 computer programers could be hired each year in the U.S. economy using the PAT. In view of the current rapid expansion of this occupation, it is likely that this number is a substantial underestimate.

It was not possible to determine prevailing selection ratios (SR) for computer programers in the general economy. Because the total yearly number of applicants for this job in the government could not be determined, it was also impossible to estimate the government SR. This information lack is of no real consequence, however, since it is more instructive to examine utilities for a variety of selection ratios. Utilities were calculated for SR's of .05, .10, .20 . . .80. The gains in utility or productivity as computed from equation (4) are those that result when a valid procedure is introduced where previously no procedure or a totally invalid procedure has been used. The assumption that the true validity of the previous procedure is essentially zero may be valid in some cases, but in other situations the PAT would, if introduced, replace a procedure with lower but nonzero true validity. Hence, utilities were calculated assuming previous procedure true validities of .20, .30, .40 and .50, as well as .00.

Using a modification of Equation (4), utilities that would result from one year's use of the PAT for selection of new hires in the Federal Government and the economy as a whole were computed for each of the combinations of SR and previous procedure validity given above. When the previous procedure was assumed to have zero validity, its associated testing cost was also assumed to be zero; that is, it was assumed that no procedure was used and that otherwise prescreened applicants were hired randomly. When the previous procedure was assumed to have a nonzero validity, its associated cost was assumed to be the same as that of the PAT, that is, $10 per applicant. As mentioned above, average tenure for government programers was found to be 9.69 years; in the absence of other information, this tenure figure was also assumed for the private sector. $\Delta\overline{U}$/selectee per year was multiplied by 9.69 to give final $\Delta\overline{U}$/selectee. Cost of testing was charged only to the first year.

Building all of these factors into equation (4), we obtain the equation actually used in computing the utilties:

$$\Delta U = tN_s\ (r_1 - r_2)\ SD_y\ \phi/p - N_s\ (C_1 - C_2)/p,$$

where:

ΔU = the gain in productivity in dollars in using the new selection procedure for one year,

t = tenure in years of the average selectee; here 9.69,

N_s = number selected in a given year; this figure was 618 for the Federal Government and 10,210 for the U.S. economy,

r_1 = validity of the "new" procedure, here the PAT; $r_1 = .76$,

r_2 = validity of the previous procedure; r_2 ranges from zero to .50,

C_1 = per applicant cost of the new procedure, here \$10,

C_2 = per applicant cost of previous procedure, here zero or \$10.

The terms SD_y, ϕ, and p are as defined previously. The figure for SD_y was the average of the two estimates obtained in this study. Note that although this equation gives the productivity gain that results from substituting for one year the new (more valid) selection procedure for the previous procedure, these gains are not all realized the first year. They are spread out over the tenure of the new employees.

Results and Discussion

Estimation of Yearly SD_y

The two estimates of SD_y were quite similar. The mean estimated difference in dollar value of yearly job performance between programers at the 85th and 50th percentiles in job performance was \$10,871 (standard error = \$1673). The figure for the difference between the 50th and 15th percentiles was 9,955 (standard error = \$1,035). The difference of 826 dollars is roughly 8 percent of each of the estimates and is not statistically significant. Thus the hypothesis that computer programer productivity in dollars is normally distributed cannot be rejected. The distribution appears to be at least approximately normal. The average of these two estimates, \$10,413, was the SD_y figure used in the utility calculations below. This figure must be considered to be an underestimate since it applied to incumbents rather than to the applicant pool. As can be seen from the two standard errors, supervisors showed somewhat better agreement on the productivity difference between "low performing" and "average programers" than on the difference between "average" and "superior" programers.

Table 2 shows the gains in productivity in millions of dollars that would result from one year's use of the PAT to select computer programers in the Federal Government for different combinations of SR and previous procedure validity. As expected, these gains increase as SR decreases and as the validity of the previous procedure decreases. When SR is .05 and the previous procedure has no validity, use of the PAT for one year produces a productivity gain of 97.2 million dollars. At the other extreme, if SR is .80 and the procedure the PAT replaces has a validity of .50, the gain is "only" 5.6 million dollars. The figures in all cells of Table 2 are quite large--larger than most industrial-organizational psychologists would, in our judgment, have expected. These figures, of course, are for total utility. Gain per selectee for any cell in Table 2 can be computed by dividing the cell entry by 618, the assumed yearly number of selectees. For example, when SR = .20 and the previous procedure has a validity of .30, gain per selectee is $64,725. As indicated earlier, the gains shown in Table 2 are produced by one year's use of the PAT but are not all realized during the first year; they are spread out over the tenure of the new employees. Per year gains for any cell in Table 2 can be obtained by dividing the cell entry by 9.69, the average tenure of computer programers.

Table 3 shows productivity gains for the economy as a whole resulting from use of the PAT or substitution of the PAT for less valid procedures. Table 3 figures are based on the assumed yearly selection of 10,210 computer programers nationwide. Again, the figures are for the total productivity gain, but gain per selectee can be computed by dividing the cell entry by the number selected. Once mean gain per selectee is obtained, the reader can easily compute total gain for any desired number of selectees. As expected, these figures are considerably larger, exceeding one billion dollars in several cells. Althought we have no direct evidence on this point, we again judge that the productivity gains are much higher than most industrial-organizational psychologists would have suspected.

In addition to the assumptions of linearity and normality discussed earlier, the productivity gain figures in Tables 2 and 3 are based on the assumption that selection proceeds from top-scoring applicants downward until the SR has been reached. That is, these analyses assume that selection procedures are used optimally. Because of the linearity of the relation between test score and job performance, any other usage of a valid test would result in lower mean productivity levels among selectees. For example, if a cutting score were set at a point lower than that corresponding to the SR and if applicants scoring above this minimum score were then selected randomly (or selected on other non-valid procedures or considerations), productivity gains would be considerably lower than show in Tables 2 and 3. (They would, however, typically still be substantial.)

The PAT is no longer available commercially. Originally marketed by Psychological Corporation, it was later distributed by IBM as part of "package deals" to computer systems purchasers. However, this practice was dropped about 1974, and since then the PAT has not been available to most users (Note 6). This fact, however, need create no problems in terms of validity generalization. The results of this study generalize directly to other tests and subtests with the same factor structure. The three subscales of the PAT are composed of very conventional number series, figure analogies, and arithmetic reasoning items. New tests can easily be constructed that correlate 1.00, corrected for attenuation, with the PAT subtests.

It should be noted that productivity gains comparable to those shown in Tables 2 and 3 can probably be realized in other occupations, such as that of clerical worker, in which lower values will be offset by the larger numbers of selectees. Pearlman, Schmidt, and Hunter (Note 7) present extensive data on the generalizability of validity for a number of different kinds of cognitive measures (constructs) for several job families of clerical work.

There is another way to approach the question of productivity gains resulting from use of valid selection procedures. One can ask what the productivity gain would have been had the entire incumbent population been selected using the more valid procedure. As indicated earlier, the incumbent population of interest in the Federal Government numbers 18,498. As an example, suppose this population had been selected using a procedure with validity of .30 using a SR of .20. Then had the PAT been used instead, the productivity gain would have been approximately 1.2 billion dollars [9.69 (18,498)(.76-.30) 10,413 (.28/.20)]. Expanding this example to the economy as a whole, the productivity gain that would have resulted is 10.78 billion dollars.

Obviously, there are many other such examples that can be worked out, and we encourage readers to ask their own questions and derive their own answers. However, virtually regardless of the question, the answer always seems to include the conclusion that it does make a difference—an important practical difference--how people are selected. We conclude that the implications of valid selection procedures for workforce productivity are much greater than most of us have realized in the past.

REFERENCE NOTES

1. Brogden, Personal communication, 1967.
2. Howard, R.A. Decision analysis: applied decision theory presented at the Fourth International Conference on Operational Research. Boston, 1966.
3. Matheson, J.E. Decision analysis practice: Examples and insights. In OR 69: Proceedings of the Fifth International Conference on Operational Research. Venice: Tavistock Publications, 1969.
4. Rosenberg, I. Gast, Schmidt, F.L., & Hunter, J.E. Application of the Schmidt-Hunter validity generalization model to computer programers. Personnel Research and Development Center, U.S. Civil Service Commission, 1978.
5. Hughes, J.L., & McNamara, W.J. Manual for the revised programer Aptitude Test. New YOrk: The Psychological Corporation, 1959.
6. Dyer, P. Personnel communication, April 1978.
7. Pearlman, K., Schmidt, F.L., & Hunter, J.E. Test of a new model of validity generalization: Results for predictors of proficiency in clerical work. Personnel Research and Development Center, U.S. Civil Service Commission, 1978.

REFERENCES

Brogden, H.E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 1946, 37, 65-76.

Brogden, H.E When testing pays off. Personnel Psychology, 1949, 2, 171-183.

Brogden, H.E., & Taylor, E.K. The dollar criterion: Applying the cost accounting concept to criterion construction. Personnel Psychology, 1950a, 3, 133-154.

Brogden, H.E., & Taylor, E.K. A theory and classification of criterion bias. Educational and Psychological Measurement, 1950b, 10, 159-186.

Cattell, R.B. Psychological measurement: ipsative, normative, and interactive. Psychological Review, 1944, 51, 292-303.

Cronbach, L.J., & Gleser, G.C. Psychological tests and personnel decisions. Urbana, Illinois: University of Illinois Press, 1965.

Curtis, E.W., & Alf, E.F. Validity, predictive efficiency, and practical significance of selection tests. Journal of Applied Psychology, 1969, 53, 327-337.

Dunnette, M.D., & Borman, W.C. Personnel selection and classification systems. Annual Review of Psychology, 1979, in press.

Ghiselli, E.E., & Kahneman, Daniel. Validity and non-linear heteroscedastic models. Personnel Psychology, 1962, 15, 1-11.

Ghiselli, E.E. Dr. Ghiselli's comments on Dr. Tupes' note. Personnel Psychology, 1964, 17, 61-63.

Ghiselli, E.E. The validity of occupational aptitude tests. New York: Wiley, 1966.

Guion, R.M. Personnel Testing. New York: McGraw-Hill, 1965.

Hawk, J. Linearity of criterion-GATB aptitude relationships. Measurement and evaluation in guidance, 1970, 2, 249-251.

Howard, R.A. Proceedings of the fourth international conference on operational research. New York: Wiley, 1966.

Howard, R.A., Matheson, J.E., & North, D.W. The decision to seed hurricanes. Science, 1972, 176, 1191-1202.

Hunter, J.E., & Schmidt, F.L. Fitting people to jobs: The impact of personnel selection on national productivity. In Human Performance and Productivity, ed. E.A. Fleishman, 1979, in press.

McNamara, W.J., & Hughes, J.L. A review of research on the selection of computer programers. Personnel Psychology, 1961, 14, 39-51.

Raiffa, H. Decision analysis: Introductory lectures on choices under uncertainty. Reading, Massachusetts: Addison-Wesley, 1968.

Roche, U.F. The Cronbach-Gleser utility function in fixed treatment employee selection. Unpublished doctoral dissertation. Southern Illinois University, 1961. (Portions reproduced in L.J. Cronbach & G.C. Gleser, Psychological Tests and Personnel Decisions, Urbana, Illinois, University of Illinois Press, 1965, pp. 254-266.)

Schmidt, F.L., Hunter, J.E., & Urry, V.W. Statistical power in criterion-related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

Schmidt, F.L., & Hoffman, B. Empirical comparison of three methods of assessing the utility of a selection device. Journal of Industrial and Organizational Psychology, 1973, 1, 13-22.

Schmidt, F.L., & Hunter, J.E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Shane, G.S. Further tests of the Schmidt-Hunter validity generalization model. Personnel Psychology, in press.

Sevier, Francis A.C. Testing the assumptions underlying multiple regression. Journal of Experimental Education, 1957, 25, 323-330.

Surgent, L.V. The use of aptitude tests in the selection of radio tube mounters. Psychological Monographs, 1947, 61, 1-40.

Thorndike, R.L. Personnel Selection. New York: Wiley, 1949, 169-176.

Tiffin, J., & Vincent, N.L. Comparison of empirical and theoretical expectancies. Personnel Psychology, 1960, 13, 59-64.

Tupes, E.C. A note on "validity and non-linear heteroscedastic models." Personnel Psychology, 1964, 17, 59-61.

U.S. Bureau of the Census. Census of Population: 1970. Subject Reports Final Report PC (2) - 7A (Table 1).

Van Naerson, R.F. Selectie van chauffeurs. Gronigen: Wolters, 1963. Portions translated in L.J. Cronbach and G.C. Gleser. Psychological tests and personnel decisions. Urbana, Illinois. University of Illinois Press, pp. 273-290.

Wolins, L. Responsibility for raw data. American Psychologist, 1962, 17, 657-658.

TABLE 1

Questionnaire

Estimation of Selection Utility

Computer Programers (GS-334)

Name______________________________Dept.__________________ Agency____________

INSTRUCTIONS

The dollar utility estimates we are asking you to make are critical in estimating the relative dollar value to the government of different selection methods. In answering these questions, you will have to make some very difficult judgments. We realize they are difficult and that they are judgments or estimates. You will have to ponder for some time before giving each estimate, and there is probably no way you can be absolutely certain your estimate is accurate when you do reach a decision. But keep in mind three things:

(1) The alternative to estimates of this kind is application of cost accounting procedures to the evaluation of job performance. Such applications are usually prohibitively expensive. And in the end, they produce only imperfect estimates, like this estimation procedure.

(2) Your estimates will be averaged in with those of other supervisors of computer programers. Thus errors produced by too high and too low estimates will tend to be averaged out, providing more accurate final estimates.

(3) The decisions that must be made about selection methods do not require that all estimates be accurate down to the last dollar. Substantially accurate estimates will lead to the same decisions as perfectly accuarte estimates.

Based on your experience with agency programers, we would like for you to estimate the yearly value to your agency of the products and services produced by the average GS 9-11 computer programer. Consider the quality and quantity of output typical of the average programer and the value of this output. In placing an overall dollar value on this output, it may help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of the average GS 9-11 computer programer at
______________________ dollars per year.

We would now like for you to consider the "superior" programer. Let us define a superior performer as programer who is at the 85th percentile. That is, his or her performance is better than that of 85% of his or her fellow GS 9-11 programers, and only 15% turn in better performances. Consider the quality and quantity of the output typical of the superior programer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value of a
superior GS 9-11 computer programer to be
______________________ dollars per year.

Finally, we would like you to consider the "low performing" computer programer. Let us define a low performing programer as one who is at the 15th percentile. That is, 85% of all GS 9-11 computer programers turn in performances better than the low performing programer, and only 15% turn in worse performances. Consider the quality and quantity of the output typical of the low performing programer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these producst and services.

Based on my experience, I estimate the value to my
agency of the low performing GS 9-11 computer programer
at ______________________ dollars per year.

TABLE 2

Estimated Productivity Increase from One Year's Use of the PAT to Select Computer Programers in the Federal Government (In Millions of Dollars)

	True Validity of Previous Procedure				
SR	.00	.20	.30	.40	.50
.05	97.2	71.7	58.9	46.1	33.3
.10	82.8	60.1	50.1	39.2	28.3
.20	66.0	48.6	40.0	31.3	22.6
.30	54.7	40.3	33.1	25.9	18.7
.40	45.6	34.6	27.6	21.6	15.6
.50	37.6	27.7	22.8	17.8	12.9
.60	30.4	22.4	18.4	14.4	10.4
.70	23.4	17.2	14.1	11.1	8.0
.80	16.5	12.2	10.0	7.8	5.6

TABLE 3

Estimated Productivity Increase from One Year's Use of PAT to Select Computer Programers in U.S. Economy

(In Millions of Dollars)

	True Validity of Previous Procedure				
SR	.00	.20	.30	.40	.50
.05	1605	1184	973	761	550
.10	1367	1008	828	648	468
.20	1091	804	661	517	373
.30	903	666	547	428	309
.40	753	555	455	356	257
.50	622	459	376	295	213
.60	501	370	304	238	172
.70	387	285	234	183	132
.80	273	201	165	129	93

706

JOB PERFORMANCE OF USAF BYPASSED SPECIALISTS

Captain William H. Cummings and Captain David S. Vaughan
USAF Occupational Measurement Center

An Apprentice Knowledge Test (AKT) is a 65-item multiple choice examination designed to measure job knowledge at the three-skill level in a particular Air Force enlisted job specialty. An AKT is also, to a large extent, a source of headaches for the Occupational Measurement Center, where the tests are written. The reason for this is some of the complaints the Center has received about the results of these tests. These complaints center on the observation that some airmen who pass the tests -- and who are therefore selected for entry into the career field at the apprentice level -- cannot do the work that is expected or required of them. This paper deals with our attempts to solve this problem and relieve some of the headaches.

One major responsibility of the Occupational Measurement Center is to construct the tests that support the Weighted Airman Promotion System, and related tests. The promotion testing program, which includes the Specialty Knowledge Tests and the Promotion Fitness Examinations, has proceeded smoothly since its inception in 1969. However, as noted the Apprentice Knowledge Test program (which currently includes 151 of the "related Tests") has run into a number of problems. These problems spring largely from some of the uses to which the tests are put, which are quite different from the uses of the promotion tests.

As noted, AKTs are used to select airmen for entry into a career field at the apprentice level, or the three-skill level. In this capacity, a major use of the AKT is to allow an airman to bypass technical school by showing his/her proficiency on the test. For example, if an airman has prior civilian or military training as an orderly, he/she may take the AKT for the Medical Services career field; or if he/she is trained in electronics, he/she can take an AKT appropriate to one of the numerous electronics career fields. If he/she passes, he/she will go directly to his/her first assignment with his/her three-skill level, rather than through technical school. This program obviously represents a major time savings for the airman and a major dollar savings for the Air Force.

To pass the AKT, the airman must score higher than thirty percent of the airmen who have previously taken the AKT. This scoring system

767

we feel, is responsible for most of the complaints with the AKT system. If most of the airmen who have already taken the test are well-qualified, then the examinee will need a substantial amount of job knowledge to pass his AKT. If not, he/she may get by with only a minimal display of knowledge. Therefore, passing an AKT does not necessarily have any meaning relative to the knowledge required to perform at the three-skill level on the job.

We have recently been developing a criterion-referenced scoring system for the AKT. Under this system, each AKT, as it is developed, would be first administered at the three-level technical school for that career field. The AKT passing score would then be established relative to the performances of recent technical school graduates on the test. Technical School graduates are a good reference group, since they are generally assumed to have the minimum knowledge required for adequate performance at the three-skill level. By passing the AKT, the bypass candidate will be demonstrating something more than good performance relative to a group of examinees with unknown qualities. He/she will be demonstrating that he/she has at least the minimum knowledge required for successful performance as an entry-level airman. Vaughan (1976a, b) has previously investigated this criterion-referencing system and found it to be a workable procedure.

Present plans call for the passing score to be set at the tenth percentile of technical school graduates' scores on the test. Use of the tenth percentile as the passing point is fairly arbitrary but based on some sound logical considerations. The percentile for the passing score should not be set extremely high, since this would require the bypassed specialists to know more than a substantial percentage of technical school graduates. This would be both unfair to the bypass candidate and wasteful in terms of training already-qualified airmen. However, the percentile should not be set at the lowest level of performance by the technical school graduates, either. Extremely low scores are likely to contain considerable error (Lord & Novick, 1978) and may reflect less job knowledge than the examinee actually has. Therefore, the tenth percentile was selected.

The present study was designed to eliminate some of the arbitrariness involved in use of the tenth percentile as the passing point. Two groups of airmen -- technical school graduates and bypassed specialists -- who had recently entered the medical services career field were compared on a job performance survey measure. This comparison provides the information needed to allow us to set the AKT passing point realistically. If the bypassed specialists perform about as well as the graduates, then a pass/fail point equal to or below that of the tenth percentile would be appropriate. If the bypassed specialists do not perform as well, a more stringent criterion may be necessary.

Another asset of this study is that it will demonstrate the extent to which criterion referencing will affect the job performances of

708

bypassed specialists. The new scoring system will, in most cases, affect the amount of basic knowledge the bypassed specialist will bring to his/her first assignment. It is not yet clear, however, how this knowledge difference will affect actual job performance.

Method

Subjects

Lists of recent three-skill-level technical school graduates and bypassed specialists in the medical services career field were obtained from the Air Force Military Personnel Center. From these lists, 306 airmen were selected for participation. Seventy-nine technical school graduates and 36 bypassed specialists returned booklets in complete and usable form, representing an overall usable return rate of 38%.

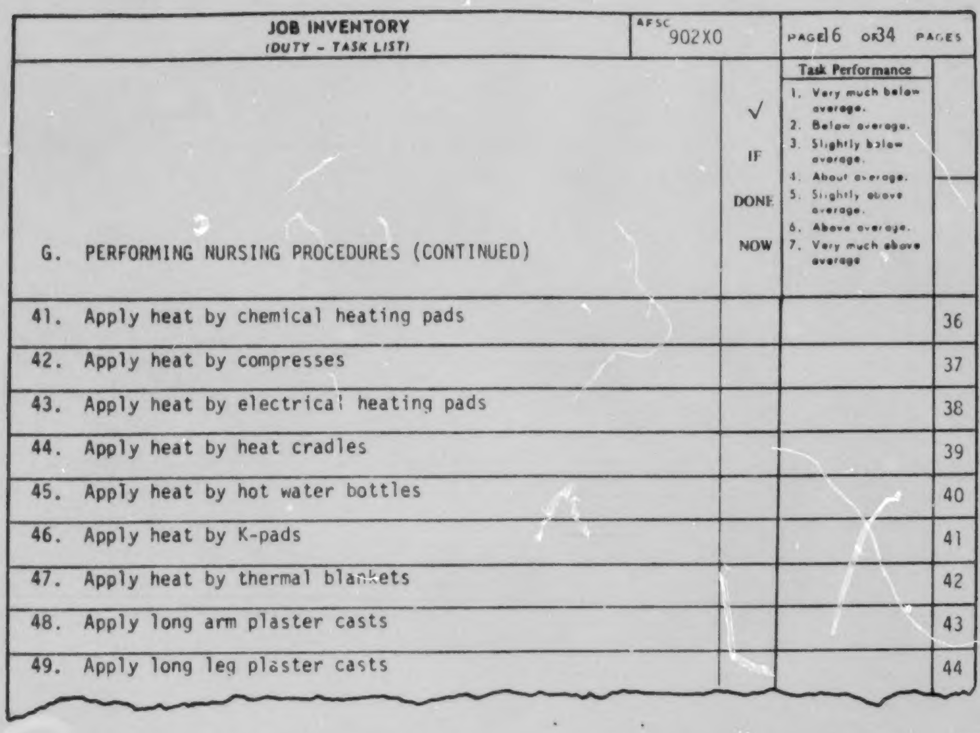
Materials

The main part of the survey materials was a modified job inventory booklet. The inventory booklets are developed by the Center's Occupational Survey Branch for the various airman specialties. Each inventory booklet is designed to contain a comprehensive list of all of the tasks that might be performed by any individual in a given specialty. Each job task very specifically describes a corresponding job behavior (e.g., "Assemble equipment for cardiac monitoring," "Administer eye irrigations," "Obtain blood from blood bank"). In the case of Medical Services, there are 505 listed tasks, and additional space is provided for up to 69 write-in tasks. In addition to the task data section, an extensive section was included for background information on the airman: historical data, time in present job, time in career field, duty area, types of equipment used, etc. A similar background information booklet provided for background data on the supervisor.

Procedure

For the purposes of this study, the survey procedure followed three steps. First, the airman was asked to complete the survey booklet by checking all of the tasks that he performed in his present job (see Fig. 1). Second, the supervisor was asked to rate the airman's performance on a 7-point scale for each task that was checked off. A sample rating scale was provided at the top of each page, indicating that a "1" represented "Very Much Below Average" performance, up to a "7", which represented "Very Much Above Average" performance. Third, approximately one month after the survey booklet was returned, the supervisor was mailed a follow-up questionnaire. This questionnaire requested a single rating of the airman's overall job performance on a 30-point Likert-type scale.

769



JOB INVENTORY (DUTY – TASK LIST)	AFSC 902X0	PAGE 16 OF 34 PAGES	
G. PERFORMING NURSING PROCEDURES (CONTINUED)	✓ IF DONE NOW	Task Performance 1. Very much below average. 2. Below average. 3. Slightly below average. 4. About average. 5. Slightly above average. 6. Above average. 7. Very much above average	
41. Apply heat by chemical heating pads			36
42. Apply heat by compresses			37
43. Apply heat by electrical heating pads			38
44. Apply heat by heat cradles			39
45. Apply heat by hot water bottles			40
46. Apply heat by K-pads			41
47. Apply heat by thermal blankets			42
48. Apply long arm plaster casts			43
49. Apply long leg plaster casts			44

Figure 1. Sample from survey booklet

Dependent Measures

Job performance data from the 505 job tasks, in addition to data previously obtained on these tasks, were condensed into four dependent measures:

TOTAL TASKS, the total number of job tasks performed by the airman. This measure was a count of all of the tasks for which the supervisor gave the airman a rating. Thus, it was not a single count of tasks the airman claimed to perform, but an indication of the tasks that the supervisor recognized the airman as performing.

$\overline{X}$ (DIFF x RATING), the average of the task performance ratings, with each task performance rating multiplied by the difficulty of that task. The task difficulty data were obtained from the Occupational Survey Report (OSR) previously available for this career field (Ballentine & Cole, 1975).

770

Job incumbents had rated the "Task Learning Difficulty" of each task ("the need for lengthy, systematic training before a new member of the appropriate Air Force Specialty could perform the task adequately") on a 1-9 scale, with a rating of "1" indicating "Least Difficult to Learn" and a rating of "9" indicating "Most Difficult to Learn". This measure provided an average measure of the airman's job performance, as opposed to TOTAL TASKS, which was a summation measure.

EQUIP ITEMS, the total number of equipment items the airman indicated that he/she used on his/her present job.

FOLLOW-UP, the airman's overall job performance, as rated on the 30-point follow-up survey scale.

Results and Discussion

Differences in Job Performance

Table 1 presents the t-tests and summary statistics comparing the two groups, and Figs. 2-5 present histograms for both groups. Neither Table 1, nor any of the histograms show any significant differences in the mean performance levels of the two groups. The mean and median performance levels of the bypassed specialists are slightly higher than those of the tech school graduates for all four measures. However, the column of t-tests shows that none of these differences is significant. Bypassed specialists show significantly more variation in number of equipment items used ($\chi^2 = 4.18$, $p < .05$, by Bartlett's test). This trend is repeated, although nonsignificantly, in the case of TOTAL TASKS and $\overline{X}$ (DIFF x RATING) but reversed in the case of FOLLOW-UP. Similarly, the histograms reflect no substantial differences in terms of either central tendency or significant numbers of outliers at the lower ends of the distributions. These analyses indicate that the bypassed specialists are at least capable of holding their own against the tech school graduates in their first job assignments, if not slightly outperforming them.

Further analyses were conducted to control for the effects of various background variables. One rather disturbing finding was the large number of airmen (73% of the bypassed specialists and 58% of the tech school graduates) who had already advanced to the five-skill level. Thus, the groups were split by current skill level, and a 2 X 2 analyses of variance (Career Field Entry Method X Current Skill Level) was performed on each measure. For $\overline{X}$ (DIFF x RATING), EQUIP ITEMS, and FOLLOW-UP, there were no significant main effects or interactions (all Fs $<$ 2, with df = 1, 92). For TOTAL TASKS, the bypassed specialists ($\overline{X} = 115.56$) tended to outperform the tech school graduates ($\overline{X} = 99.93$); $F(1, 92) = 3.49$, $p = .06$. The main effect of Skill Level ($F = 1.62$) and the interaction ($F = 1.83$) were non-significant. Similarly, analyses of covariance, which incorporated various background variables as covariates, failed to reveal any significant or

Table 1

Comparative Job Performance Data

Variable	Group	$\overline{X}$	Mdn	SD	t
		Measure			
TOTAL TASKS	Tech School	104.16	103.50	42.70	- 1.29
	Bypass	116.28	112.50	53.60	
$\overline{X}$ (D x R)	Tech School	22.60	23.25	4.72	- 0.84
	Bypass	23.43	23.42	5.20	
EQUIP ITEMS	Tech School	16.57	16.17	7.30	- 0.63
	Bypass	17.61	17.10	9.75	
FOLLOW-UP	Tech School	23.13	25.59	6.19	- 0.85
	Bypass	24.21	25.75	5.77	

noteworthy differences between the two groups. These analyses consistently demonstrate that, at least for this specialty, the bypassed specialists perform up to the level of the technical school graduates.

Effects of Criterion Referencing on the 902X0 Career Field

An additional question of considerable interest is the degree to which the criterion referencing procedure will affect the actual job performances of bypassed specialists in this career field. This specialty was one of those examined in the earlier criterion referencing studies (Vaughan, 1976b). Previous tech school graduates had taken the same AKT that the bypassed specialists in this study took, and the raw score corresponding to the tenth percentile of the tech school graduates' performances was computed. This is the passing score that would be established under the proposed criterion-referenced scoring system. In this case, that raw score was 34 (out of a possible total score of 65 points). The actual passing score for this AKT varied between 26 and 32 points (depending on the performance of previous AKT candidates). Thus, there is a cluster of bypassed specialists within this sample who passed the AKT but who would have failed under the new system. These "theoretical failures" are plotted in Figs. 2-5 as the small arrows above each histogram for the bypass group.

It is obvious that the theoretical failures are quite evenly scattered throughout the distributions and that criterion referencing would have very little effect on the job performances of bypassed specialists in this

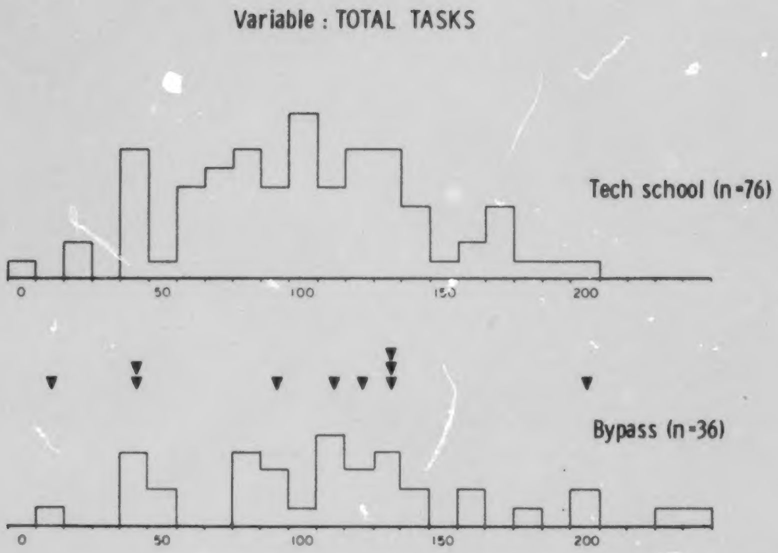


Figure 2. Distributions of TOTAL TASKS scores.

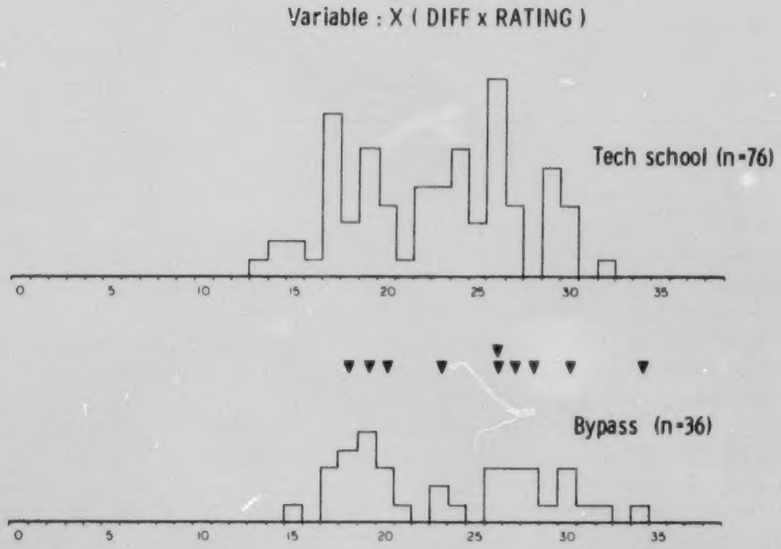


Figure 3. Distributions of X (DIFF x RATING) scores.

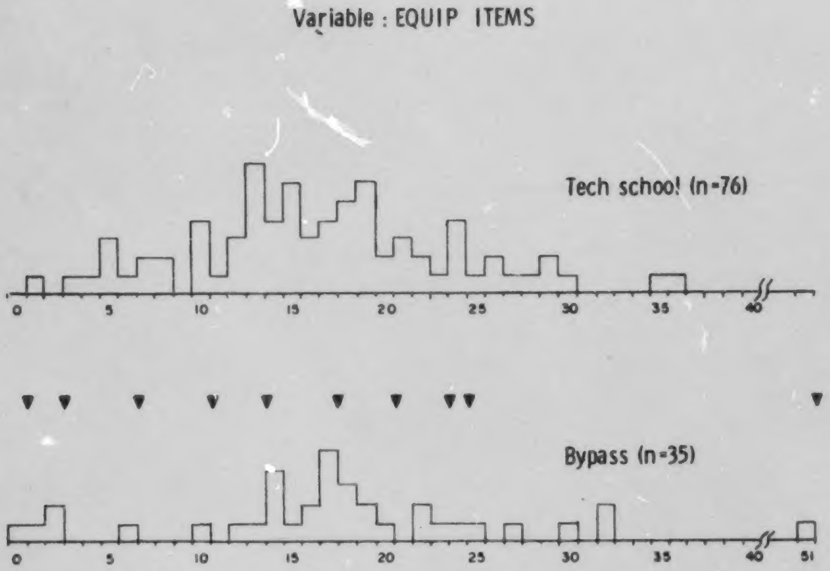


Figure 4. Distributions of EQUIP ITEMS scores.

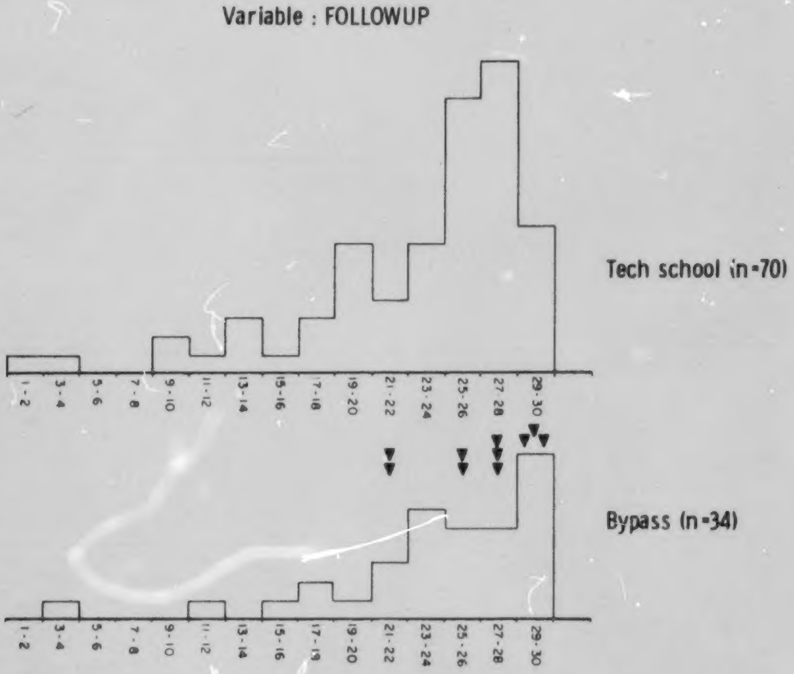


Figure 5. Distributions of FOLLOWUP scores.

career field. However, it does not seem appropriate at this point to conclude, in general, that criterion referencing will not have any effect on job performance. The Medical Services career field is fairly unique in two aspects. First, the theoretical passing score was fairly close to the actual passing score (ranging from two to eight points away). Second, there is a great deal of pre-service experience available in this career field (e.g., orderly, nurse's aide). It is likely that most AKT examinees in this area, who claim relevant job experience or job knowledge are more-or-less qualified to do the work. Exact placement of the criterion may not be too important in this type of specialty. The case may be quite different for other career fields. This state of affairs indicates the need to replicate this study in other career fields.

Criterion-Related Validation of the AKT

This study provides a unique opportunity to validate one of the Center's tests against certain job performance measures. If, in fact, the AKT does not correlate with at least some of these measures, this finding in itself would have important implications for use of the AKT and positioning of the pass/fail criterion.

The correlation coefficients between AKT scores and the performance measures indicate that the AKT did _not_ correlate significantly with $\overline{X}$ (DIFF x RATING), r (34) = .12; EQUIP ITEMS, r (34) = .03; or FOLLOW-UP, r (32) = .03. However, the AKT _did_ correlate significantly with TOTAL TASKS, r (34)= .33, $p < .05$. This pattern of correlations indicates that AKT scores do not predict the airman's average performance level, but they do predict what, or how many different things, the airman can do. Therefore, the AKT does appear to be a good screening instrument for determining award of the three-skill level.

Conclusions

The major finding of this study is that the bypassed specialists did about as well as -- even slightly better than -- the technical school graduates on all performance measures. This finding holds for both the raw measures and the measures corrected for various background variables. The implications of these results for the major question of this study -- where to set the AKT passing criterion -- is quite clear: The criterion should be set _no higher_ than the tenth percentile of technical school graduates' scores on the AKT. A higher cutoff would only tend to block the flow of qualified three-level airmen into the career field.

The question of the effect of criterion referencing on the job performance levels of bypassed specialists is less easily answered. Certainly the higher pass/fail point would have little effect in this career field. However, it is not clear that this finding will generalize to other career fields. We anticipate that the effect may be quite different in specialties where pre-service experience is less readily available or where the criterion-referenced passing score is farther from the current passing score. This issue can only be resolved by additional job performance studies of bypassed specialties.

We feel that the Occupational Survey-based job performance survey technique will prove to be of considerable value not only in terms of bypassed specialist performance but in a variety of other situations as well. However, our immediate concerns are with further performance studies of the bypassed specialist population. As the criterion-referencing system goes into effect, it will become important to extend these findings to other career fields. The present data indicate no substantial differences between bypassed specialists and technical school graduates. If these findings generalize to other career fields, we can accept the tenth percentile of technical school graduates' AKT scores as the passing point for the AKT with considerable confidence.

References

Ballentine, R. D., & Cole, G. B. Occupational survey report of medical service career ladder (AFPT 90-902-191). Lackland Air Force Base, TX.: Occupational Survey Branch, USAF Occupational Measurement Center, December 1975.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Vaughan, D. S. Criterion-referencing the 47330 Apprentice Knowledge Test to technical school graduates. Lackland Air Force Base, TX.: USAF Occupational Measurement Center Technical Note Series, March 1976. (No. 76-01)

Vaughan, D. S. Development of criterion-referenced Apprentice Knowledge Tests. Lackland Air Force Base, TX.: USAF Occupational Measurement Center Technical Note Series, May 1976. (No. 76-04)

Acknowledgements

The authors would like to gratefully acknowledge the many people who assisted with this project: Col James Turner, Commander, USAF Occupational Measurement Center; Mr. Stephen Fotis, Chief, Occupational Test Development Branch; Dr. Walter Driskill, Chief, Occupational Survey Branch; Lt Col Russell Johnson; Major Stan Stephenson; Capt David Street; 1Lt Mike McMillan; 2Lt Linda Alford; Mr. Henk Ruck; SSgt Bill Snyder; Sgt Bill Fogarasi; and Airman Wendy Metzinger.

ARRO
ADVANCED
RESEARCH
RESOURCES
ORGANIZATION

4330 East West Highway [illegible] • [illegible]

ANALYSIS OF
HEAVY EQUIPMENT OPERATOR JOBS*

Sidney A. Fine
Howard C. Olson
David C. Myers
Margarette C. Jennings

A. Michael Collins--Project Coordinator for IUOE

*Presentation to Military Testing Association
Airport Hilton Inn West
Oklahoma City, Oklahoma
November 1, 1978

a division of RESPONSE ANALYSIS CORP [illegible]

ANALYSIS OF HEAVY EQUIPMENT OPERATOR JOBS

I. PROBLEMS AND OBJECTIVES

This report concerns a project known as the TRAINING STANDARDS PROJECT (TSP), conducted for the past two-and-a-half years under the auspices and with the active participation of the International Union of Operating Engineers (IUOE).* The union, in collaboration with contractors, conducts a national apprentice training program at some 75 training centers operated by local unions throughout the United States.

During the early 1970's, class action suits initiated by individuals were brought against several IUOE locals charging racial discrimination in selection for apprenticeship training. Among the charges were that the required high school diplomas, language and mathematics requirements of qualification tests, and the length (4 years) of apprenticeship were either irrelevant to the work or unnecessary to achieve competence. The work of the operating engineer, it was charged, was much simpler than was claimed by the union, was so classified in the Dictionary of Occupational Titles of the United States Employment Service, and could be learned in a much shorter period of time.

In view of the inadequacy of data to deal with these charges, the union sought and obtained support for research to establish for itself and for the public the true nature of the work and the skill required. In undertaking this research, it not only was concerned with the response to the courts and affirmative action in the area of equal employment opportunity, but also with the improvement of its own training practices. Included in its objectives for the Training Standards Project were:

- To define the work of the operating engineer so that the knowledge, skills and abilities required could be satisfactorily communicated to the courts and the public.
- To establish training standards for every important operating engineering task.
- To provide a basis for more objective and defensible apprentice selection procedures--namely, tests.

---

*A. Michael Collins, the union's present monitor, worked very closely with ARRO personnel, particularly in arranging and managing the active participation and contribution of union members.

778

In the initial planning of the project around these objectives, a number of needs emerged that dictated the choice of methodology and procedure. They were as follows:

- The job analysis data developed needed to satisfy the courts as to the level of knowledges, skills, and abilities required, but also needed to
- *define performance standards*, and
- *training required to meet the standards*, and take into account
- *regional and environmental differences in performance*. Finally, the analysis had to produce
- *content valid measures* that did not result in discrimination against particular groups of people in our society.

This report will focus on the work done to satisfy these needs and meet the union's objectives.

II. TECHNICAL APPROACH

Job Analysis Phase. The union employed S. A. Fine Associates to design, manage and carry out the research, a decision made to some degree because of their interest in the Functional Job Analysis (FJA)* methodology used by this organization. FJA focuses on tasks which are formulated as fundamental, stable units consisting of a behavior and a result. These tasks are organized into job assignments in one combination or another to accomplish a job of work. Data for preparing task statements are obtained in observation/interviews. In addition to the behavior and result, the task statement includes information about the resources the worker draws upon--the machine tools and equipment used and the level of instructions, that is, the prescription/discretion mix that the worker must follow. The accuracy and reliability of the task statement are controlled by 10 ratings on ordinal scales functionally defined that establish the level of complexity with regard to Things, Data, People, Instructions, Reasoning, Mathematics, and Language. From this information it is possible to directly formulate Performance Standards and Training Requirements. The complete task analysis provides the information to fulfill the paradigm: To do this task to these standards, the worker needs this training.

---

*S. A. Fine and W. W. Wiley. *An introduction to Functional Job Analysis: A scaling of selected tasks from the welfare field*, No. 4. The W. E. Upjohn Institute for Employment Research, September 1971.

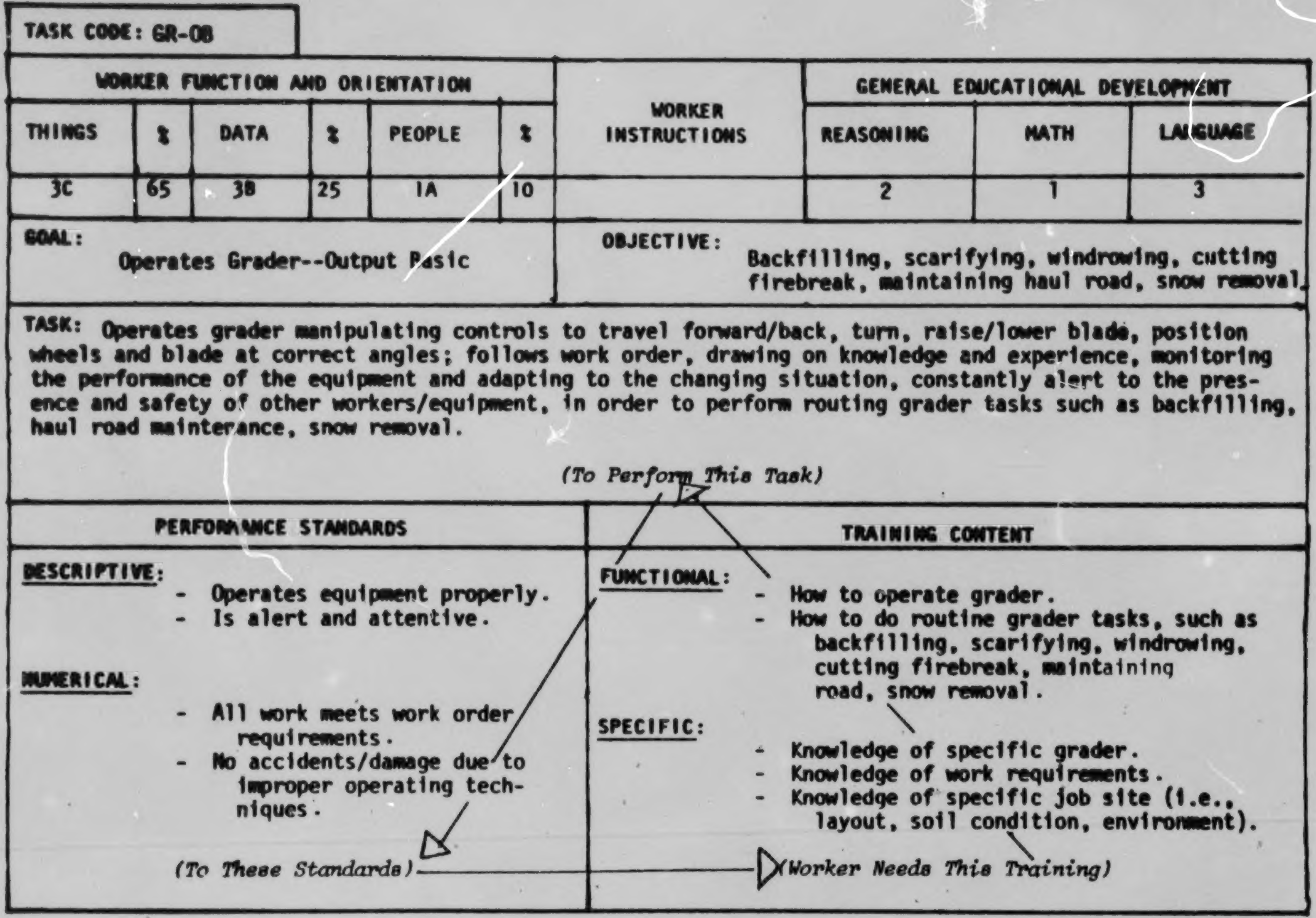
Functional Job Analysis was used to develop baseline information about operator requirements necessary to perform the tasks and produce the necessary outputs that are within the capability of a piece of equipment. Job analysis was conducted for 16 kinds of construction equipment normally operated by operating engineers. This paper deals only with the bulldozer, backhoe, loader, grader, and scraper, in the so-called blade category of equipment. To illustrate,a completed task statement for the Grader is shown in Fig. 1.

A cadre of some 20 senior operating engineers, engaged in apprenticeship training and experienced across the full range of the jobs being analyzed, received a week of training in Functional Job Analysis methods. They, then, took on assignments as individuals to serve as expert consultants to the consulting psychologists on one or another piece of equipment. Through them arrangements were made to visit training and job sites where observation/interviews were conducted. Subsequently, in task force groups of 3 or 4, they reviewed the various *drafts* of the task analysis for accuracy, coverage, and communicability. Task statements were then edited and made consistent in form by the consulting psychologist. The final step in the job analysis phase was the assembly of the total group of 20 FJA-trained operators where the assembly as a whole reached consensus on what should be included for each item of equipment.

Performance Standards Phase. Performance standards for the operation of a piece of heavy equipment are intended to describe the jobs that an experienced operator should be capable of performing with that machine. The standards are cast in terms of specific outputs (types of results) and operator behaviors required to accomplish each output safely, efficiently, and effectively.

The job analysis for an item of equipment is usually represented in seven task statements (seven printed pages):

- Inspects the equipment (prior to operation)
- Services the equipment
- Starts the equipment
- Operates the equipment--basic outputs
- Operates the equipment--intermediate outputs
- Operates the equipment--difficult outputs
- Shuts down the equipment.



TASK CODE: GR-08

WORKER FUNCTION AND ORIENTATION						WORKER INSTRUCTIONS	GENERAL EDUCATIONAL DEVELOPMENT		
THINGS	%	DATA	%	PEOPLE	%		REASONING	MATH	LANGUAGE
3C	65	3B	25	1A	10		2	1	3

GOAL: Operates Grader--Output Basic

OBJECTIVE: Backfilling, scarifying, windrowing, cutting firebreak, maintaining haul road, snow removal.

TASK: Operates grader manipulating controls to travel forward/back, turn, raise/lower blade, position wheels and blade at correct angles; follows work order, drawing on knowledge and experience, monitoring the performance of the equipment and adapting to the changing situation, constantly alert to the presence and safety of other workers/equipment, in order to perform routing grader tasks such as backfilling, haul road mainterance, snow removal.

(To Perform This Task)

PERFORMANCE STANDARDS

DESCRIPTIVE:

- Operates equipment properly.
- Is alert and attentive.

NUMERICAL:

- All work meets work order requirements.
- No accidents/damage due to improper operating techniques.

(To These Standards)

TRAINING CONTENT

FUNCTIONAL:

- How to operate grader.
- How to do routine grader tasks, such as backfilling, scarifying, windrowing, cutting firebreak, maintaining road, snow removal.

SPECIFIC:

- Knowledge of specific grader.
- Knowledge of work requirements.
- Knowledge of specific job site (i.e., layout, soil condition, environment).

(Worker Needs This Training)

Figure 1. Illustrated Task Statement for the grader

The performance standards are detailed expansions of the standards listed in the task statements, exploring behavioral implementations of various contingencies as well as critical "know-how" developed through experience. They run from 100 to 150 pages for each piece of equipment.

The standards are stated in terms of those that are primarily mental in nature, requiring planning, monitoring, and checking; those that require interpersonal relationships; and those that require the combination of perception and physical coordination to accomplish the operations, such as manipulating controls and operating the equipment to meet work specifications. For example, the backhoe output of "precision excavating"* is described in 24 mental/planning/monitoring standards, 7 interpersonal standards, and 43 physical action standards.

The process for developing performance standards comprised four steps:

- The psychological consultant prepared a preliminary draft of the performance standards to establish a common format among all standards.
- The preliminary draft was reviewed during a two-day meeting of the consultant and a subject matter expert for that piece of equipment.
- The standards were revised, incorporating changes decided on in the previous step, and resubmitted to the subject matter expert for approval.
- The proposed standards were reviewed and revised by a Task Force selected for that piece of equipment. In the "task force" review meeting (requiring two days), each output was discussed, one-by-one, and decision reached as to proper wording and description of each performance within the output.

Performance standards task forces of 4 to 6 subject matter experts were formed for each piece of equipment in the project. Task force members were selected to be geographically representative of operating engineers nationwide, so variations in operating practice as a function of climate, region of the nation, equipment model preferences, and so on, are taken into account.

Test Development Phase. Work sample performance tests were developed directly from the performance standards. Those tasks that were most often performed were chosen on the advice of the task force subject matter experts.

---

*The fully qualified backhoe operator should be able to perform eight outputs (in addition to the common outputs of inspection/servicing, and start-up/shut-down: (1) compacting with a vibratory attachment, (2) loading a haul vehicle, (3) removing trees and stumps, (4) pavement breaking, (5) filling and backfilling, (6) hoisting, (7) placing riprap, and (8) precision excavating.

Most test layouts consisted of a set of formalized work requirements. The operator being examined received instructions much in the same way that a job foreman would issue them. The operator then read the grade stakes and performed the earth moving necessary to meet the job specifications. Test situations varied from 1 1/2 to 3 hours. The tasks making up the work samples are shown in Table 1.

Performance on the work sample was timed, and measured by a series of test items drawn nearly verbatim from the performance standards. Items are statements covering three general areas of equipment operation: skill in operating the equipment, the safety behavior and practices demonstrated, and the extent to which the job specifications were met. The items are arranged in checklist format with space for (a) simple Yes/No checks to indicate whether or not the behavior was observed, and (b) ratings on a 5-point scale (1 = poor, 5 = superior) for overall performance and the satisfying of task specifications. There are generally about 20 items of each type for each output of the test. Each test was tried out before use to assure the correct time allocation, sufficiency of the instructions, appropriateness of the tasks performed, and the ease with which the test could be used.

Test Administration. Validation of the work sample performance tests has been conducted with the same care and attention to detail exercised in the development of performance standards. The locations of each of the week-long validation testings is shown in Table 2. Again test sites were chosen to be geographically dispersed across the nation. The bulldozer test validation was the first conducted with testing of 28-32 operators at each of four locations. It later was decided that more subject operators could be tested at fewer locations without sacrifice to the integrity of the validation. For all subsequent validation testing, 36-42 operators were tested at each location. In all, 360 operators made up the validation sample.

The validation strategy was to test operators of prejudged, known skill levels, with test administration and scoring by subject matter experts (in the study, the test administrators are called "observers") who have no knowledge of the prejudged, known skill levels of the operators. Then, if a test differentiates as presumed, the most highly skilled operators will perform better on the test than the average, who in turn will perform better than the least skilled. It is obvious that the operator selection beforehand is the key to a "valid" validation.

TABLE 1

Outputs Tested for Each Piece of Equipment

Bulldozer (3 hours)	Excavate for foundation, backfill
	Finish a slope
	Pushload scraper, run fill
	Cut and fill, build ramp
	Build bench
Backhoe (2 hours)	Excavate vertical wall trench
	Expose buried pipe
	Excavate sloping wall trench
	Excavate pier hole
Loader (1 1/2 hours)	Excavate basement
	Form spoil pile
	Load haul vehicle from stockpile
Grader (2 1/2 hours)	Build maintenance road
	Cut rough ditches
	Level material and crown road
	Construct V-ditch to grade
	Finish grade to a flat surface
Scraper (Varied)	Load scraper
	Haul material to fill area
	Unload scraper
	Return to cut area

TABLE 2

Locations and Dates of Validation Testing

Equipment	Date	Place
Bulldozer	May 1977	Cleveland, Ohio
Bulldozer	May 1977	Philadelphia, Pa.
Bulldozer	May 1977	Des Moines, Iowa
Bulldozer	May 1977	Sacramento, Calif.
Backhoe	May 1978	Dayton, N. J.
Loader	June 1978	Beaumont, Calif.
Loader	July 1978	New Alexandria, Pa.
Grader	June 1978	Columbus, Ohio
Scraper	July 1978	Richmondville, N. Y.
Scraper	August 1978	Seattle, Wash.

The idealized quota of subject operators at each location was 32 operators for bulldozer testing and 40 operators for the other kinds of equipment, distributed as follows:

	Bulldozer	Other Equipment
Level 3 Journeyman Operators	8	10
Level 2 Journeyman Operators	8	10
Level 1 Journeyman Operators	8	10
Senior or Newly Graduated Apprentices	8	10
	32	40

The description of each category is shown in Table 3. In addition to the quota by skill level, operator selection committees were instructed to select, if possible, without compromising skill level representation, 30 percent minority.* Minority representation, in actuality, turned out to be 90 out of 360 or 25 percent.

Operator selection committees were formed at each test location. Committee members included Operating Engineer Apprenticeship Coordinators, some contractors, and business agents, and dispatchers for the local union. The skill category classification of the operators selected to participate was held strictly confidential, with no one other than the selection committee and psychologist in charge knowing these classifications.

III. RESULTS

The data analysis and results for each test focused on the following important issues:

> Does the test do what it is supposed to do? Is it valid?
>
> In considering recent EEOC Guidelines, what influence does certain operator characteristics, specifically race, have on test performance and validity?

The analysis was designed to determine whether the tests in whole and in part differentiate between the four criterion groups; and if racial membership creates differences in test performance and test validity.

To address these issues and to answer related questions, three areas were explored: (1) the statistical organization of the tests, (2) the validity of each test, and (3) the differences between white and minority operators' test performance.

---

*As distinguished from the nonminority, "White," operators, the Minority operators include: Black (origins in Black African racial groups), Hispanic (Mexican, Puerto Rican, Cuban, Central or South American or other Spanish culture), Asian or Pacific Islander, American Indian, and Alaskan Native.

TABLE 3

Operator Selection Criteria

- Level 3 Journeyman--This class of operator is the most expert of all on that particular piece of equipment and is considered to be a "top hand." Such an operator can work to specifications essentially on his or her own, and usually can perform all of the outputs of which the machine is capable; most significantly, the Level 3 Operator's work is likely to never need follow-up by another operator.

- Level 2 Journeyman--This class of operator is the broad class of "average" operators. These operators may be skillful in some outputs, but never in all. This class of operator usually can manage on his or her own without much supervision. However, the Level 2 operator's work occasionally will need follow-up by a more skilled operator.

- Level 1 Journeyman--This class of operator may not have had experience on all of the outputs that a machine is capable of performing, or may, despite experience, lack the skills to perform the outputs well; the operator needs a lot of supervision. What is most characteristic is that the operator's work often will not meet, exactly, performance criteria or output specifications; most significantly, the Level 1 operator's work will often need follow-up by a more skilled operator.

- Apprentice--The apprentice could fall within any of the three journeyman skill levels. Most likely, however, since apprentices' experience usually is limited, the apprentice skill level will be at Journeyman Level 1 or lower. All that is required for the performance checklist validation is that the apprentices participating (a) be in their third or fourth years of apprenticeship training (or recent graduates), and (b) have had some training on the piece of equipment being tested.

Organization of the Performance Tests. To search for commonality among test items, a principal components analysis was performed on three of the performance tests (i.e., loader, backhoe, and grader). The analyses were followed by several orthogonal rotations and an oblique rotation. Of these, the Varimax method provided the most simplified and meaningful factorial structures.

With some slight variation, the analyses yielded three components of heavy equipment operations that were relatively stable across the three pieces of equipment. The components were: using correct procedure and meeting specifications, operating with caution and safety, and following instructions. For example, in using the correct procedures and meeting specifications, the expert loader operator manipulates controls with precision and smoothness, travels forward into material, tilts bucket to aid breakout, fills bucket without straining the engine, empties bucket at 45 degree angle, and obtains uniform grade. The expert operator also functions with caution and safety, such as, not abusing equipment and following safety rules. To follow instructions, he/she comprehends instructions from supervision and interprets grade stakes properly.

Criterion-related Validity. The item analysis for each of the five performance tests included the usual statistics, such as, means, standard deviations, and frequency distributions. The discrimination index was also calculated for each item to determine if the item indicated differences between levels of operator skill.

Table 4 illustrates an item from the grader test that significantly differentiated between the four criterion groups. In other words, the operators who rated high on the item (i.e., the operator manipulated the controls with precision and smoothness) were in fact the *participants* identified by the selection committee as expert and above average operators, while the participants who were rated lower on the item had been classified as below average operators and senior apprentices.

The item analyses indicated that, overall, about 80 percent of the test items significantly differentiated between the levels of operator skill. The results demonstrate that a majority of the test contents are valid and therefore indicative of the operator's level of competence.

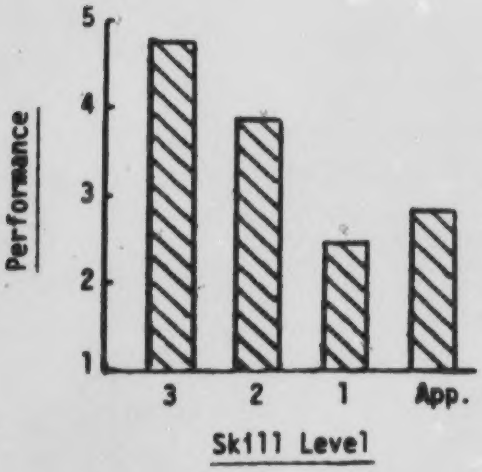
TABLE 4

Example of a Discriminating Question

Question: Did the grader operator manipulate controls with precision and smoothness?

		Observer Ratings					Total No. of Oper.	Aver. Rating
		1	2	3	4	5		
Skill Level	3	-	-	1	1	11	13	4.77
	2	1	-	1	3	3	8	3.87
	1	3	3	5	2	-	13	2.46
	App.	-	2	2	1	-	5	2.80
Totals		4	5	9	7	14	39	

p = .001 (highly statistically significant)



Since the overall rating of operator performance was significantly correlated with the remaining items in each test, it was used to demonstrate the overall validity of the tests. Figure 2 clearly demonstrates the significant relationships between test performance and skill level.

Of further importance, findings for the loader performance test demonstrate the uniqueness of loader operations. The principal components analysis yielded three dimensions similar to the other equipment (e.g., grader and backhoe); however, in contrast to other equipment, efficiency of performance and economy of effort (e.g., operator avoids excessive turning and traveling) emerged as salient aspects of loader operations.

During the development of the loader performance standards and tests, it became evident that an effective indicator of efficiency is the operator's cycle time (i.e., the period from when the operator initially tilts the bucket to empty the material until he/she begins to tilt the bucket to dump the next load of materials). Because of its critical nature, cycle time was included as an item and recorded for each operator during the test. Figure 3 illustrates again the validity of the loader test. Those participants who had previously been classified as below average operators and apprentices had cycle times of about 60 seconds, while the participants who were previously classified as expert operators had significantly shorter cycle times of about 37 seconds. The expert loader operators demonstrated greater efficiency and economy of effort in their performance during the test than the below average operators and apprentices.

To summarize, the performance-based tests for the five pieces of equipment are doing what they were intended to do, i.e., differentiating between levels of operator skill. The results demonstrate criterion-related validity.

Differences in Test Performance and Validity for White Operators and Minority Operators. The next phase of the analysis involved separating the total sample of participants for all pieces of equipment into white operators (N = 290) and minority operators (N = 70). Figure 4 illustrates the results.

Even though there are slight differences in test performance between white and minority operators at each skill level, the important point illustrated by the figure is that as skill level increases for both white and

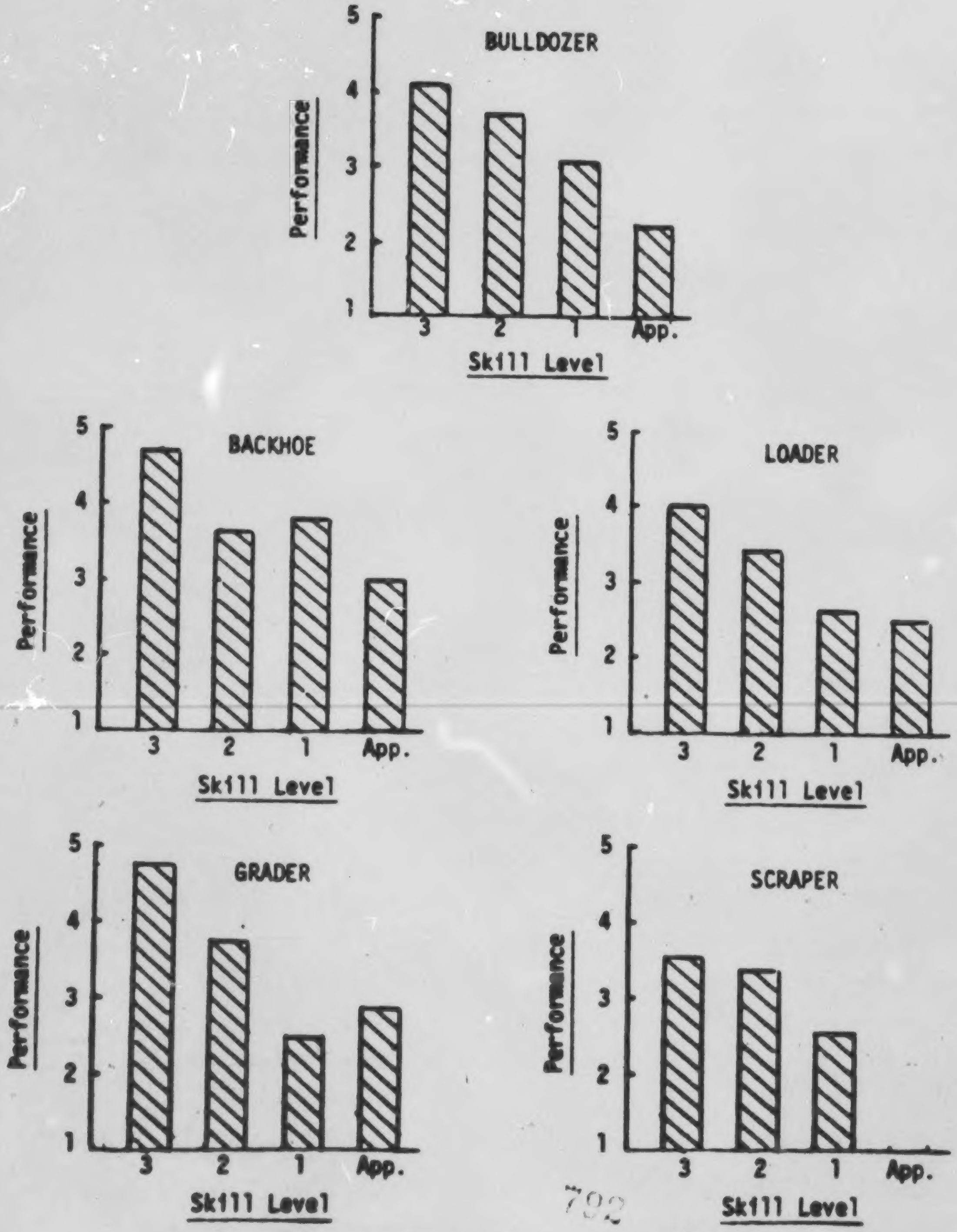


Figure 2. Performance vs. skill level by equipment

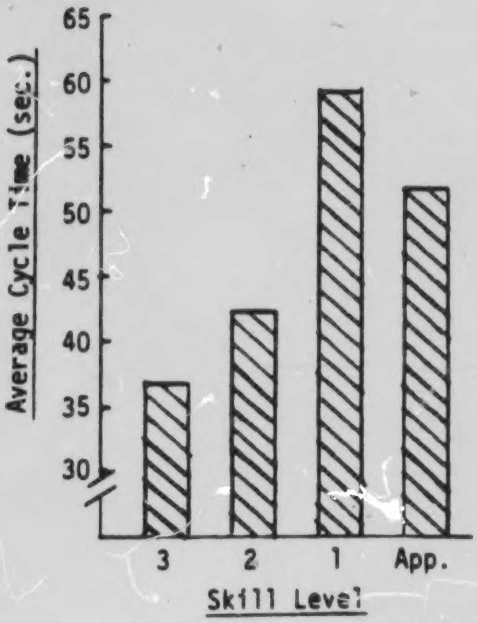


Figure 3. Loader cycle time

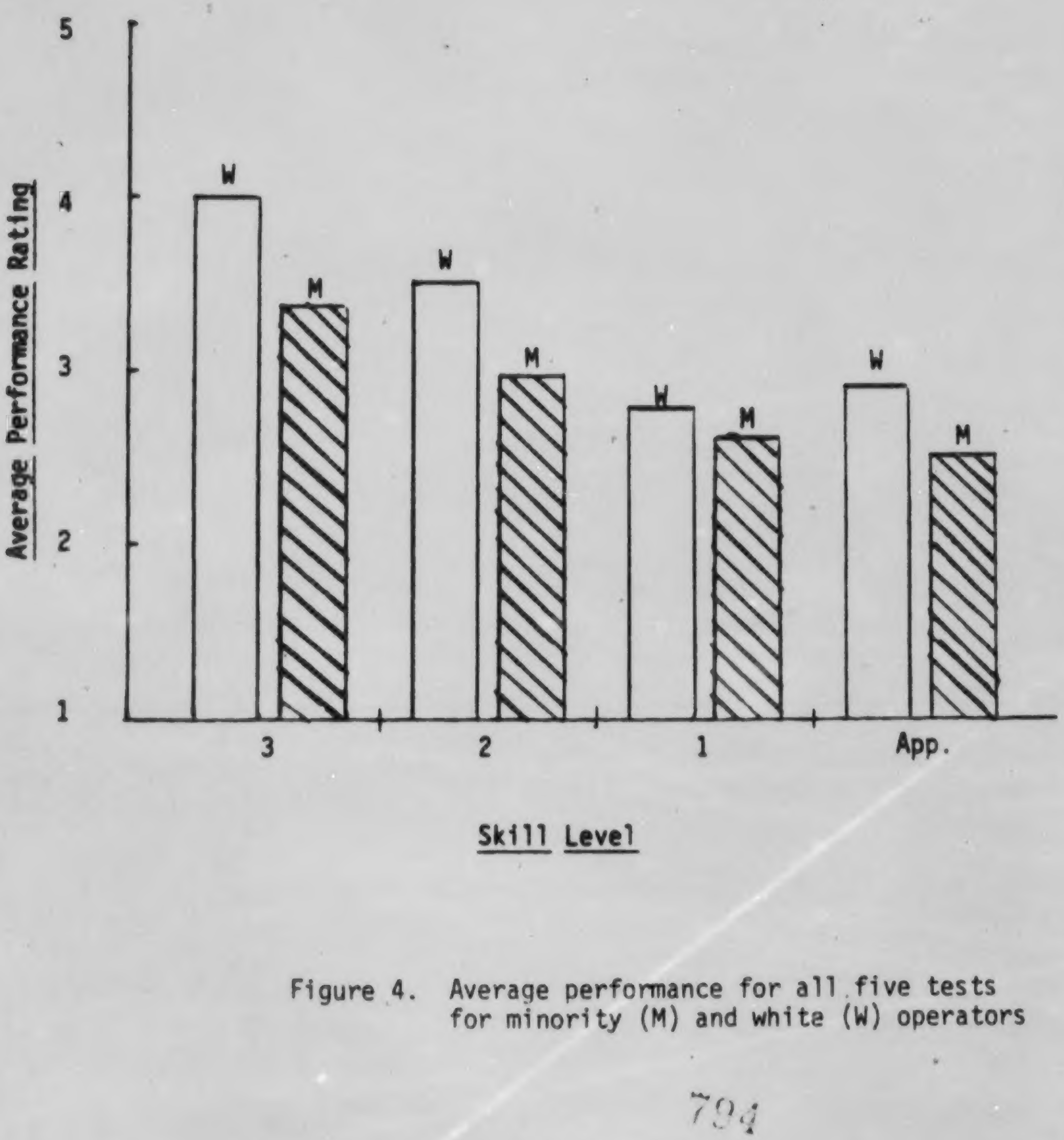


Figure 4. Average performance for all five tests for minority (M) and white (W) operators

794

minority operators, so does their test performance. In other words, the tests are valid for both groups of operators, although at somewhat different levels of performance.

To further investigate the differences in test performance between the white and minority groups, we looked at differences in education, age, and experience. In contrast with age and education, the difference between the number of years of experience in heavy equipment operations between white and minority participants was highly significant. It appears that these differences in experience between whites and minorities partially explains the differences in test performance. In other words, the minority operators have less experience with heavy equipment operations than do white operators. Consequently, as a group, they have acquired fewer of the necessary operating skills and thus perform less well on the tests.

In summary, while there were differences in test performance between white and minority operators, the test performance for both racial groups increased significantly as their skill level increased. The tests are valid for both white and minority operators.

IV. CONCLUSIONS

The conclusions we have drawn from the data presented can be grouped under three headings--Content, Criterion, and Construct validity--topics of central concern in the EEOC Guidelines.

Content Validity. The tests are content valid. The items on the tests are performance based, reflect representative outputs for each piece of equipment, and have been selected by subject matter experts of the craft. Although not yet tested in the courts, there seems little doubt that the tests will meet the guideline requirements as fair measures for qualifying apprentices as journeymen.

The performance tests proved to be a positive, satisfying learning experience for observers and operators at all levels. A typical post-test operator reaction was: "What a great way to check myself out." Observers, several of whom were instructors, felt the tests could serve as effective checklists for union instructors. Because of these kinds of reactions and the demonstrated validity, the union has begun the production of slide/tape training films for apprentices using the same performance standards as the basic source for script and associated curriculum materials.

Criterion Validity. Using the "known group" technique for establishing criterion groups, in a concurrent validity frame of reference, it is clear that the tests have criterion validity. By and large, the tests significantly discriminate between skill levels among operating engineers in which the significant criterion factor is the degree of independence and autonomy the operator can be permitted in doing the work. This is of vital importance in a trade in which the operator is entrusted with extremely expensive machines and charged with accomplishment of work basic to both the success of a construction project and the safety of many other workers.

Construct Validity. The factor analyses indicate the possibility of three constructs pertinent to the functional and specific content of heavy equipment operation. The constructs are the three components yielded by principal components analysis, namely:

- Using correct machine operating procedures and meeting output specifications.
- Following instructions (these are in good measure operator initiated).
- Operating with caution and safety.

The three constructs correlate closely with the Things, Data, People categories for the generation of the performance standards.

It is noteworthy that the data and people relationships of operating engineers, usually overlooked in standard descriptions, were uncovered and articulated by the use of Functional Job Analysis, as significantly involved in their work. The Things, Data, People aspects of the components need to be further researched as a basis for developing aptitude tests for entry applicants.

PREDICTIVE UTILITY OF THE OFFICER EVALUATION BATTERY (OEB)

by

Arthur C. F. Gilbert, Ph.D.

A Paper Prepared for Presentation at the
20th Annual Conference of the Military Testing Association (MTA)
Oklahoma City, Oklahoma
October 30 - November 3, 1978

Performance and Training Research Laboratory
U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

797

PREDICTIVE UTILITY OF THE OFFICER EVALUATION BATTERY (OEB)

Arthur C. F. Gilbert, Ph.D.
U. S. Army Research Institute for the
Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

The Officer Evaluation Battery (OEB) was designed to measure a number of dimensions predictive of success as an Army officer. The Officer Evaluation Battery is essentially the same test battery as the Cadet Evaluation Battery (CEB), the development of which is discussed by Mohr and Rumsey (1978). The only difference between the two instruments is in terms of purpose of administration. The CEB is administered to cadets in the Army Reserve Officers Training Corps (ROTC) for selection and/or counseling purposes while the OEB is for administration to newly commissioned officers for experimental purposes.

The Officer Evaluation Battery consists of cognitive and non-cognitive subtests. The seven subtests are Combat Leadership (Cognitive), Technical Managerial Leadership (Cognitive), Career Potential (Cognitive), Combat Leadership (Non-Cognitive), Technical Managerial Leadership (Non-Cognitive), Career Potential (Non-Cognitive), and Career Intent (a non-cognitive scale). The seven subtests of the OEB and the types of items in each are shown in Table 1. Earlier research (Helme, Willemin, and Grafton, 1974) indicated the utility of the OEB item content in predicting success in a simulated combat situation.

The purpose of this research was to evaluate the predictive utility of the OEB in Officer Basic Courses (OBC). An Officer Basic Course exists for each of the 13 Career Branches in the Army. A newly commissioned officer attends one of these courses on entering upon active duty prior to his initial duty assignment. The research focussed on evaluating the predictive effectiveness of each of the subtests and the combination of subtests in relation to Officer Basic Course final grades for the total sample and within the three different types of Officer Basic Courses: Combat Arms, Combat Support, and Service Support. Another purpose of the research was to determine if there were differences in prediction for males and for females and to evaluate possible differences in prediction for black officers and for white officers.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the view of the U. S. Army Research Institute or the Department of the Army.

798

Table 1

Officer Evaluation Battery (OEB) Subtests and Description of Items

SUBTEST	DESCRIPTION OF ITEMS
Combat Leadership (Cognitive)	Military tactics; practical skills in a variety of areas ranging from out-door activities to mechanical and electronic applications.
Technical-Managerial Leadership (Cognitive)	History, politics; culture; mathematics; physical sciences
Career Potential (Cognitive)	Technological knowledge relevant to military requirements.
Combat Leadership (Non-Cognitive)	Combat leader qualities, occupational interests, sports interest, outdoor interests related to combat leadership
Technical-Managerial Leadership (Non-Cognitive)	Mathematics and physical sciences skills and interest; urban or rural background; scientific interest and ability; decisive leader qualities; and verbal-social leadership
Career Potential	Clerical-administrative interest, versus white collar interest, combat interest
Career Intent	Intention of making the Army a career choice

Procedure

The Officer Evaluation Battery was administered to all officers in the 13 Career Branches who attended Officer Basic Courses during Fiscal Year 1974. Final OBC course grades (i.e., the criterion measure) were collected from each OBC for as many subjects as possible.

The initial sample consisted of 9,180 officers but this sample included many officers who entered on active duty for training only and who did not enter a duty assignment on completion of the Officer Basic Course. Since it was felt desirable to keep the validation sample as homogeneous as possible, only those officers who continued on in an active duty status after Officer Basic Course were retained in the validation sample. A total of 4,622 were so identified. However, for some of these officers, complete data (i.e. OEB scores and OBC final course grades) were not available.

Final course grades were reported by the different schools as either percentage grades or as class standings within the OBC; in some cases both percentage grades and class standings were reported. When only grades were reported, they were rank ordered and a class standing generated for each student. The resulting class standing were converted to Army standard scores. Where class standing was available, it was converted directly to an Army standard score.

A multiple regression analysis was performed using all seven subtests of the OEB as predictors with final OBC grades as the criterion for the total sample. The intercorrelation matrix was computed using pairwise deletion of missing values in the data matrix. The total sample was divided on the basis of type of Officer Basic Course (i.e., Combat Arms, Combat Support, and Service Support). The total sample was also divided on the basis of sex and finally on the basis of race (i.e., black officers and white officers). Since these breakdowns could only be accomplished where classification data were available, the number in the subgroups will not always equal the total number of cases. Parallel analyses were then performed for each of the subgroups (i.e., seven separate analyses).

Results and Discussion

The correlations between each of the OEB subtests and Officer Basic Course final course grades are shown in Table 2 for the total sample. The multiple correlation of all seven subtest scores with the criterion are also presented. The same data are presented for each of the seven analyses in this table.

800

Table 2

Correlations Between Each Officer Evaluation Battery (OEB) Subtest and Officer Basic Course Final Grades for the Total Sample and for Each Subsample

OEB Subtest	Total Sample (N=2,836)	Combat Arms (N=1,536)	Combat Support (N= 903)	Service Support (N= 397)	Male Sample (N=2,719)	Female Sample (N= 113)	Black Sample (N= 190)	White Sample (N=2,603)
Combat Leadership (Cognitive)	.36**	.31**	.43**	.33**	.36**	.42**	.41**	.29**
Technical Managerial (Cognitive)	.29**	.20**	.29**	.37**	.29**	.33**	.29**	.22**
Career Potential (Cognitive)	.32**	.26**	.40**	.36**	.32**	.33**	.27**	.26**
Combat Leadership (Non-Cognitive)	.16**	.19**	.14**	.01	.16**	-.01	.27**	.14**
Technical Managerial (Non-Cognitive)	.28**	.17**	.19**	.17**	.22**	.22**	.23**	.17**
Career Potential (Non-Cognitive)	.12**	.17**	.08**	-.08	.13**	-.15	.22**	.07*
Career Intent	.09**	.15**	.03	-.08	.09**	.16	.16**	.10**
Multiple Correlation	.42**	.38**	.49**	.47**	.41**	.55**	.49**	.34**

* Significant at the .05 level.
** Significant at the .01 level.

All zero order correlations between each of the subtest scores and OBC final course grades were significant at the .01 level with the exception of the Career Intent subtest which yielded a low positive correlation of .09 with the criterion that was significant only at the .05 level. The multiple correlation of the seven subtest scores with the criterion of .42 was significant at the .01 level for the total sample.

In the Combat Arms branches, all of the seven subtests yielded zero order correlations with the criterion that were significant at the .01 level and a multiple correlation of .38. Six of the subtests yielded zero order correlations significant at the .01 level in the Combat Support branches; the only exception was the correlation between the Career Intent subtests and the criterion. A multiple correlation of .49 was obtained between the seven subtests and the criterion for this sample. All three of the OEB cognitive subtests were significantly correlated with the criterion in the Service Support branches (at the .01 level) as well as the Technical Managerial (Non-Cognitive) subtest. Small negative or negligible correlations were derived for the remaining three scales with the criterion. The multiple correlation for this subsample was .47.

A test for the significance of differences was performed among the sets of zero order correlations for the three types of branches as well as test of significance of the difference among the multiple correlations. The three cognitive scales of the OEB yielded significantly lower correlations (p less than .01) in the Combat Arms branches than in the Combat Support branches. There was not any significant difference between the Combat Support branches and the Service Support branches in terms of the predictive effectiveness of these three subtests. The Career Potential (Non-Cognitive) scale yielded a higher (significant at the .05 level) zero order correlation with the criterion for the Combat Arms branches than for the Combat Support branches. The Career Intent subtest also yielded a significantly greater correlation with the criterion in the Combat Arms branches than in the Combat Support branches; the difference was significant at the .01 level. The multiple correlation of .38 for the Combat Arms branches was significantly lower, at the .01 level, than the multiple correlation obtained in the Combat Support branches.

For the male sample all seven subtests yielded zero order correlations with the criterion that were significant at the .01 level. The multiple correlation was .41 for this sample. In the female sample, four subtests, the three cognitive subtests and the Technical Managerial (Non-Cognitive) subtest, yielded zero order correlations with the criterion that were significant at the .01 level. The corresponding multiple correlation for the sample was .55. There were not any significant differences between the zero order correlations for the two samples and there was not any difference between the two multiple correlation coefficients.

The zero order correlations between each of the OEB subtests and the criterion, as well as the resulting multiple correlation coefficient of .49, were all significantly different from zero at the .01 level for the sample of black officers. For the sample of white officers all of the zero order correlations with the criterion were significant at the .01 level with the exception of the Career Potential (Non-Cognitive) subtest that yielded a correlation of .07 with the criterion that was significant at the .05 level. The resulting multiple correlation of .34 for the sample of white officers was significant at the .01 level.

The results of this research indicate that the Officer Evaluation Battery (OEB) is a useful predictor of final course grades in the Officers Basic course. Some fluctuations occur in the different samples but these are probably a function of varying sample sizes and sample characteristics. Generally, the OEB appears to have utility in predicting the performance of junior officers in acquiring skills and knowledges necessary in their performance as Army officers.

References

Helme, W. H., Willemin, L. P. and Grafton, F. C. Prediction of officer behavior in a simulated combat situation (Research Report 1182). Alexandria, VA.; U. S. Army Research Institute for the Behavioral and Social Sciences, March 1974 (NTIS Nr. AD 779 445)

Mohr, E. S. and Rumsey, M. G. Cadet Evaluation Battery: A Comparison of 1975 Male and Female Scores with One Another and with 1971 Male Scores (Technical Paper 330). Alexandria, VA.; U. S. Army Research Institute for the Behavioral and Social Sciences, September 1978.

ASSESSMENT CENTER VARIABLES AS PREDICTORS OF ON-JOB PERFORMANCE CHARACTERISTICS

Charles H. Cory, Ph.D.

Navy Personnel Research and Development Center
San Diego, California 92152

Paper presented at

The 20th Annual Military Testing Association Conference

Oklahoma City, Oklahoma

30 October - 3 November 1978

---

[1] The opinions or assertions contained herein are those of the writer and are not to be construed as official or reflecting the views of the Navy Department.

ASSESSMENT CENTER VARIABLES AS PREDICTORS OF ON-JOB PERFORMANCE CHARACTERISTICS

A set of eight hands-on tests and a semi-structured interview administered in an assessment center was developed by Siegel and Wiesen to supplement the ASVAB for selection and assignment of General Detail personnel in the Navy. After an experimental administration of the battery to 140 male enlisted personnel, follow-up was carried out in the Fleet to validate the tests against four supervisory ratings of on-job performance.

In general assessment center variables when used separately had about the same predictiveness for job performance criteria as the ASVAB, or when used in conjunction with ASVAB increased the shrunken multiple regression coefficients from .05 to .32. The tests which were useful for supplementing ASVAB were the semi-structured interview and measures of attentional time-sharing and coordinative speed and accuracy. Shrunken multiple validity coefficients of batteries composed of the five most predictive operational and assessment center variables ranged from .38 to .75 for the four supervisory ratings of on-job performance.

Attenuation of the validity coefficients for the unreliability in supervisors' marks substantially increased the shrunken multiple regression coefficients. The findings suggest that when compensation is made for the substantial inherent unreliability in supervisors' marks, predictive validities of optimally selected batteries of written tests and assessment center variables account for most of the reliable variance in supervisors' ratings of GDPs.

Paper-and-pencil tests used for selection of personnel for unskilled and semi-skilled labor and for trades jobs have frequently been found to have low predictive validities for criteria of on-job performance. Anderson, Rousch, and McClary (1973) in a study of coil winders found that none of the paper-and-pencil tests of the GATB correlated significantly either with supervisors' ratings of overall performance or with production records. Navy studies (Cory, 1976a, 1976b, and Cory, Neffson, Rimland, & Thomas, 1978) have generally found validity coefficients of paper-and-pencil tests in the personnel classification battery with supervisor's ratings of on-job performance for unskilled and semi-skilled types of positions which ranged from .15 to .20. Maximum shrunken multiple correlations of the set of classification tests for these positions with on-job performance generally have been found to range between .20 and .25.

Ghiselli, in The Validity of Occupational Aptitude Tests (1966), a comprehensive survey, reported average validity coefficients of paper-and-pencil measures of the types used in the ASVAB with performance proficiency in skilled trades jobs which ranged from .18 to .26. He also reported correlations with these criteria of .20 to .25, on the average, for measures of finger and hand dexterity and of .29 for measures of personality.

When the Ghiselli data were reclassified into categories of Skilled, Semi-skilled and Unskilled jobs, additional interesting relationships were found. Thus the average validity coefficients of paper-and-pencil tests generally decreased from Skilled to Semi-skilled to Unskilled jobs, except for measures of Perceptual Speed, where the direction of the relationship was reversed. Personality tests also had higher average validities for Semi-skilled and Unskilled jobs than for Skilled jobs, but the coefficients of Finger and Hand Dexterity measures were about the same for the three types of jobs. The obvious conclusion from Ghiselli's findings is that broadening the set of predictors to include measures of coordination and dexterity and of personality together with the paper-and-pencil measures of the ASVAB is likely to improve the Navy's ability to select and classify personnel for unskilled and apprenticeship types of jobs.

For this reason, Siegel and Wiesen (1977) developed for the Navy a battery of tests to be used to assign the personnel who are not sent to Navy Technical schools. These are roughly the individuals in the bottom 25 percent of enlisted personnel in terms of mental ability. These individuals are usually assigned as General Detail personnel (GDPs) to commands. There they work in manual labor and semi-skilled types of maintenance and housekeeping jobs with the eventual objective of training on-the-job for positions in the trades or technical areas. GDPs do not normally receive formal academic training for jobs following completion of Recruit Training and a two-week general apprenticeship training course.

The Siegel and Wiesen tests were administered in an assessment center setting which was denominated a "Technical Classification Assessment Center" after the types of jobs for which it was designed to select. The purpose of this paper is to describe the predictive validities for on-job performance of General Detail personnel of the assessment center variables in comparison with those of the classification test scores and biographical variables which were available operationally.

Data Collection

During November and December 1975, 140 male enlisted graduates from Recruit Training at the Naval Training Center in San Diego were examined at the TCAC. A description of the testing results as well as the development and characteristics of the measures in the TCAC is given in Siegel and Wiesen (op.cit.). Two separate follow-ups were carried out to collect on-job performance marks on these personnel.

Thus in November 1976 survey questionnaires were sent out to commands in the Fleet to which personnel from the Siegel and Wiesen study were currently assigned. These special questionnaires collected on-job performance marks for the men from their current supervisors. At the same time the personnel office for the command was requested to forward a record of the last set of operational performance marks that the man had received. Three months later, in February 1977, the request was repeated to commands which had not responded to the initial questionnaire. Finally, in August 1977, approximately six months after the first set of supervisors' marks was collected, a second follow-up was carried out to collect a new set of the same criterion marks.

Assessment Center Variables

Eight job performance tests and a semi-structured interview were used in the TCAC, from which 29 scores were derived. In addition four global ratings were made by the assessment center staff. The assessment center tests together with their scores are briefly described in the following slides and commentary:

Slide #1- 1. Conceptual Integration/application--a troubleshooting problem
in which a simulated hypothetical system was described together with
Slide #2- possible malfunctions and their causes. Then a series of malfunctions
was presented and the examinees were asked to identify on the basis of
the symptomatic conditions, the causes. Score was the number of correct
answers.

2. Inspection/sort--a timed test consisting of 90 items, each being one of six types. The task was to sort the items by type and to reject those which had imperfections or did not closely match the type definition. Scores computed consisted of total numbers and percentages of items correctly sorted/rejected and incorrectly sorted/rejected and an unweighted composite of the number of items correctly sorted plus the number of items correctly rejected.

3. Reliability--a variation on one of the Hartshorn and May exercises (1930) required the threading of 15 needles and self-reporting of the number of needles successfully threaded. Since the eyes of five of the needles were blocked by clear plastic and consequently could not be threaded, any score above ten was considered to be a lie. Score was a binary variable coded "1" for a truthful response and "0" for a lie.

4. Tool and Object Nomenclature, use and recognition--a test in which unusual tools and objects from Navy life were presented and briefly discussed. Subsequently three 15-item true-false tests covering the material were administered. Scores were computed measuring examinee's ability to associate an object with its (1) use and (2) name, and (3) his ability to associate its name with its use. An unweighted sum of these scores served as a fourth variable.

Slide #3- 5. Dual Task--a test designed to measure an individual's ability to
Slide #4- carry out attentional time-sharing while doing simultaneously two separate
tasks. The test required monitoring a control panel while fabricating a
pipe assembly. Cues presented on the control panel specified changes to
be made in settings of the panel. Scores were computed for the number of
parts assembled correctly and the number of panel settings performed

correctly together with the response latencies for execution of the panel settings.

Slide #5- 6. Coordinative Speed and Accuracy--a timed test using a simple wiring task. After a brief instruction and practice session was given, the examinee's task was to connect wires between terminals located on separate panels. Directions for the interconnections were shown in a Slide #6- wiring diagram and a color-coding chart. Scores were computed for the total number of connections which were correct.

7. Level of Aspiration--a dart throwing task having three sets of trials. Prior to each trial the examinee estimated the score which he would obtain for the trial. Variables scored included the candidate's estimated score for the first trial, the sum of the estimated scores for the first and the second trials and the total number of times the estimated score was lower than the score received on the previous trial (considered to be a measure of pessimism). In addition the examining staff recorded binary global marks for each candidate indicating the presence or absence of three attributes: realism, pessimism and optimism.

8. Social Interactive Evaluation--a group task covering a simulated ammunition storing problem in which members of the group transported sand in buckets to and from bins over a course which had bottlenecks and difficult transportation points. Three timed trials were interspersed by team planning sessions which were designed to critique and improve team coordination. Scores computed were algebraic sums of the positive and negative behaviors expressed by the individual in interacting with the group during each of the following: (a) the first trial, (b) the first planning session and the second trial, (c) the second planning session and the third trial, and (d) all trials and planning sessions.

9. Interview--a semi-structured interview conducted by a 2-person panel. General topics and extent of coverage, but not the individual questions, were specified for the interview. Ratings were made on 16 categories covering interest, personality characteristics, and motivation. Based on them a mean evaluation on the interview was computed and global scores were made on the candidate's ability on each of the following dimensions: (a) learning, (b) psychophysical/motor, and (c) social/motivational.

In addition, the following summary evaluations based on the total findings of the Assessment Center were made: (a) global evaluation of ability of the examinee to perform on-the-job, (b) algebraic sum of the positive and negative comments about the examinee recorded during the testing, and (c) number of discussions and votes of Assessment Center personnel required to arrive at agreement concerning evaluation decisions.

Operationally-derived Variables

A description of the operational classification tests and the biographical measures which were used in the study is shown in the next slide.

810

Slide #7- The six operational test scores used were from the personnel classification battery used by the Navy at the time. These tests were predecessors to ASVAB tests, but their areas of measurement and characteristics were very similar to those of tests in the ASVAB battery.

Although the four biographical measures shown (variables 7 to 10, inclusive) were collected specifically for the present study, the first three variables are present in Navy operational records and could be used for selection purposes, if desirable. However, because of strictures in the Privacy Act, the last variable would probably not be available to the Navy at the present time. Fortunately, however, it proved not to be useful as a predictor anyway.

Criteria

Twelve of the criteria which were used for the study are described Slide #8- on the next slide. Of these criteria, Professional Performance, Military Behavior, Military Appearance, and Adaptability describe specific aspects of behavior, or traits. Ratings for these traits were collected from two sources: (1) the operational performance ratings and (2) the ratings collected on the special questionnaire. Two global performance marks, AV-Special and AV-Op consisted of unweighted averages of the four trait marks. AV-Special was computed from the marks given on the Special Questionnaire. In contrast AV-Op was computed from the man's official performance marks. Two other global ratings, OVER and REEN, consisted of single-element marks which were collected from the special questionnaires.

Analysis

After the questionnaire returns had been merged with the records from from the TCAC, test-retest reliability coefficients were computed for supervisors' marks. Then zero-order and multiple-regression validity coefficients were computed for the four global criteria which were collected on the first follow-up. A step-wise procedure, the accretion method was used to compute multiple regression coefficients, and estimates of the shrunken validities were computed using a technique recommended by Thiel (1971).

For each criterion the predictor set used for multiple regression was restricted to those variables whose zero-order coefficients were statistically significant. Additional restrictions imposed at each step were (1) that the _F_ ratio of the incremental variation in the criterion predicted by the independent variable selected for the step with the unpredicted variation of the criterion was ≥ 4.5, and (2) that the proportion of the variance of the independent variable which was not explainable by the variables already selected was $>.30$. These restrictions were imposed in order to limit the variables selected to those which made real as opposed to chance contributions to the predictiveness of the battery.

In addition a hierarchical selection mode was employed in which variables were made available to the regression program a set at a time for the three sets of variables: (1) operational tests, (2) biographical variables, and (3) assessment center variables. The first two of these sets were composed of variables which were or which could be derived operationally, and the last set contained the variables which were being used experimentally as predictors. Thus the hierarchical mode permitted the computation and evaluation of the incremental validity added by biographical variables and by assessment center variables to the maximum validities available from the tests in the operational classification battery.

Results

At the time of the first mailout, 14 persons in the sample had been discharged or were carried as deserters. Questionnaires were returned for 106 personnel, an 85 percent return rate for the 125 who remained in the Navy at the time. For the second follow-up 31 personnel had left the service or were deserters and 71 questionnaires were returned, a 66 percent return rate. The return rate for the second follow-up undoubtedly was lowered because the usual second mailout to non-respondents was omitted in order to expedite the study.

Zero-Order Validities of Operational and Assessment Center Variables

Zero-order validity coefficients of the operational and the assessment center variables which had statistically significant coefficients de #9- for any of the four global performance marks are shown in the next slide. Twenty-eight of the 136 coefficients were statistically significant. For the operational test, biographical, and assessment center variables, respectively, 21, 19, and 15 percent of the predictors were statistically significant. Coefficients of the statistically significant variables ranged from .19 to .34 for the operational variables and from .21 to .50 for the assessment center variables. In general the lowest values were for REEN and the highest values were for AV-Op. ARI and YRED were the major operational variables which were significantly predictive of supervisors' marks. Of the ten assessment center variables which had statistically significant validities for global on-job performance, four were significant for only one criterion and six were significant for two or more criteria. Six of the nine assessment center tests had statistically significant predictive relationships with supervisors' global marks. These tests were Coordinative Speed and Accuracy, Inspection/Sort, Tool and Object Naming, Dual Task, Level of Aspiration, and Mean Interview Rating.

Maximally Predictive Sets of Operational and Experimental Variables

The shrunken, step-wise multiple correlation coefficients for maximally predictive sets of the operational, biographical, and the assess- Slide #10- ment center variables are shown in the next slide. Each row in the table represents the addition of a predictor. The total number of predictors selected for a criterion at any one point is shown by reading down the column to that point.

These data indicate that batteries formed from ARI and YRED of the operational variables had shrunken validity coefficients ranging from .33 to .42 for the four global marks. Assessment center variables added from .05 to .33 to bring about maximum shrunken validity coefficients for these criteria which ranged from .38 to .75. Thus operational variables accounted for a maximum of 11 to 18 percent of the variance of supervisors' marks. Addition of assessment center variables to this battery would increase the predictive accuracies so that from 14 to 56 percent of the variance could be predicted. The predictive accuracies for the two supervisory marks which were composites, AV-Special and AV-Op were particularly high. In an analysis which has been set forth elsewhere, it was concluded that the higher validities of the composite marks resulted, at least in part, from their greater reliability.

Differences in the types of variables selected for the maximally predictive batteries for the four criteria suggest that there were differences in the characteristics of the criteria as they were perceived by supervisors. The single-element criteria, OVER and REEN, seem to be largely focussed on professional competence. The variables which were maximally predictive for these criteria were cognitive tests, years of education, and measures of accuracy of perception and execution in hands-on situations. In contrast, the two composite marks reflected not only these characteristics, but also characteristics of personality and attitude.

Computation of Attenuated Values for the Predictive Validities

Correction of the validities of the supervisors' marks for unreliability in the criteria was also carried out in order to provide more realistic estimates of the actual predictive validities of the operational and assessment center measures. For this purpose the test-retest reliability coefficients of the eight trait and the four global marks were Slide #11- computed. These coefficients are shown in the next slide. As you may recall, the coefficients were computed by correlating each of the performance marks received for the first follow-up with the same variable from the second follow-up, which was collected approximately six months later. The statistics in the table are shown for a Total (T) Sample and for a Diverse (D) Sample, that subgroup for whom the supervisor completing the second questionnaire was different from the one completing the first questionnaire. Ninety-two percent of the personnel for whom identifying information for supervisors was available were in the D Sample. There are no entries in the D Sample for operationally-derived marks because it was considered to be not desirable to collect information identifying the supervisors completing the operational marks.

In general the test-retest reliabilities were quite low. They ranged from .16 to .58 for the trait marks and from .29 to .55 for the global marks. The reliabilities of the two global single-element marks were from .16 to .26 lower than those for the two global composite marks. AV-Op was the most reliable global mark, and REEN was the least reliable one.

The average reliability of the trait marks was .07 higher for the T Sample than for the D Sample. Although the reliability coefficients of all of the marks with counterpart values were statistically significant for the T Sample, for the D Sample two trait marks and one global mark were not significant. However, statistical tests indicate that differences between counterpart values in the T and the D Samples were not significant. Also, in general, the relative magnitudes of the reliabilities of the trait and the global marks were the same for the T and the D Samples. Therefore it was felt that the comparisons could justifiably be carried out on the sample with the greater number of degrees of freedom, the Total Sample.

Attenuated Zero-order and Multiple Regression Coefficients

Zero-order values for the validity coefficients corrected for unreliability of the supervisors' marks are shown in the next slide.

Slide #12-

In general, increases in magnitude of the zero-order coefficients caused by the attenuation ranged from .09 to .22. Some of the coefficients in the table are very high. However, the asterisks indicating the statistical significance of their values are the same as those shown for the unadjusted coefficients previously presented.

Estimates of the attenuated multiple regression coefficients for the four global marks were made by substituting the attenuated zero-order validity coefficients into the predictor-criterion intercorrelation matrix and recomputing the multiple regression statistics. For this step only variables whose unattenuated coefficients had been significant were made available to the regression program.

Slide #13- The recomputed statistics, shown in the next slide, indicate that when adjustments were made for the unreliability of the criteria, the accuracy achievable from the battery of tests and biographical variables was very high. These figures indicate that operational variables accounted for from 35 to 45 percent of the variance of supervisors' marks and that an additional 32 to 59 percent of that variance would be accounted for by assessment center variables. The total set of operational and assessment center predictors would account for from 77 to 94 percent of the reliable variance of supervisors' marks.

Although I have used conservative procedures for making these estimates, the predictiveness of the total battery of tests seems surprisingly large. However, even if the magnitude of the findings is discounted somewhat the data still show these variables to be predicting most of the reliable variance of the supervisors' marks.

In summary the major findings of the study were:

1. A composite formed from operational classification test scores and years of education accounted for between 35 and 45 percent of the reliable variance of supervisors' ratings of on-job performance of General Detail personnel. Addition of assessment center variables to this battery resulted in a total battery which accounted for 77 to 94 percent of the reliable variance of supervisors' marks.

2. The assessment center variables which were the most useful as predictors measured work accuracy under time-sharing conditions, speed and accuracy of finger-hand dexterity or coordination, classification accuracy in a hands-on situation, and personality and attitudinal characteristics.

3. Supervisors' marks formed from composites of two or more scores were more reliable than those which were based on only a single rating element.

These findings will be checked on a new sample of 1,000 to which a revised form of the TCAC has been administered. In the event the findings hold up, it is hoped that eventually the TCAC may be used to identify the incoming GDPs who have potential for advancing into a technical rating so that these personnel can be channeled into appropriate assignments before they are lost to the system.

REFERENCES

Anderson, Harry E. Jr., & Roush, S. Larry. Relationships Among Ratings, Production, Efficiency, and the General Aptitude Test Battery Scales in an Industrial Setting. *Journal of Applied Psychology*, 1973, *58*(1), 77-82.

Cory, C. H. *An Evaluation of Computerized Tests as Predictors of Job Performance: II. Differential Validity for Global and Job Element Criteria*. San Diego: Navy Personnel Research and Development Center, January 1976.

Cory, C. H. *A Comparison of the Job Performance and Attitudes of Category IVs and I-IIIs in 16 Navy Ratings*. San Diego: Navy Personnel Research and Development Center, May 1976.

Cory, C. H., Neffson, N., Rimland, B., & Thomas, E. D. *The Validity of a Battery of Experimental Tests in Predicting Performance of Navy 100,000 Personnel*. (In preparation, 1978).

Ghiselli, Edwin E. *The Validity of Occupational Aptitude Tests*. New York: John Wiley & Sons, Inc., 1966.

Hartshorne, H., May, M. A., & Shuttleworth, F. K. *Studies in the Organization of Character*. New York: MacMillan, 1930.

Siegel, A. I., & Wiesen, J. P. *Experimental Procedures for the Classification of Naval Personnel*. San Diego: Navy Personnel Research and Development Center, January 1977

SLIDE #7

VARIABLES AVAILABLE OPERATIONALLY

CLASSIFICATION TESTS

GENERAL CLASSIFICATION TEST	(GCT)
ARITHMETIC TEST	(ARI)
MECHANICAL REASONING TEST	(MECH)
CLERICAL APTITUDE TEST	(CLER)
ELECTRONIC TECHNICIAN SELECTION TEST	(ETST)
SHOP PRACTICES TEST	(SHOP)

BIOGRAPHICAL VARIABLES

YEARS OF SCHOOLING COMPLETED	(YRED)
AGE TO NEAREST BIRTHDAY	
DEMERITS IN RECRUIT TRAINING	
ARREST RECORD, BINARY CODE	

Slide #8

CRITERIA

VARIABLE	SOURCE[a]
TRAIT SCORES	
PROFESSIONAL PERFORMANCE	0
PROFESSIONAL PERFORMANCE	S
MILITARY BEHAVIOR	0
MILITARY BEHAVIOR	S
MILITARY APPEARANCE	0
MILITARY APPEARANCE	S
ADAPTABILITY	0
ADAPTABILITY	S
GLOBAL MARKS	
AVERAGE OF THE OPERATIONAL TRAITS (AV-Op)	0
AVERAGE OF THE SPECIAL TRAITS (AV-SPECIAL)	S
OVERALL PERFORMANCE (OVER)	S
RECOMMENDATION FOR REENLISTMENT (REEN)	S

[a] 0 = OPERATIONAL, S = SPECIAL QUESTIONNAIRE

SLIDE #9

ZERO-ORDER VALIDITIES FOR GLOBAL CRITERIA

VARIABLE	CRITERION[a] AV-Op	 AV-SPECIAL	 OVER	 REEN
OPERATIONAL VARIABLES				
ARI	26*	21*	30**	28*
ETST	24*	08	10	10
YRED	32**	34**	19*	18
ASSESSMENT CENTER SCORES				
PREDICTED PERFORMANCE IN RECOMMENDED ASSIGNMENT	29*	08	09	08
NUMBER OF DISCUSSIONS AND VOTES	19	21*	16	21*
NUMBER OF COMMENTS	25*	39**	17	17
INSPECTION/SORT: ACCURACY FOR DEFECTIVE ITEMS	-26*	-05	-10	-13
TOOL AND OBJECT NAMING: SCORE 1	15	09	11	21*
DUAL TASK: NUMBER OF CORRECT SETTINGS	33**	26**	23*	12
COORDINATIVE SPEED & ACCURACY: ACCURACY	18	16	25**	21*
PERCENTAGE ACCURACY	24*	28**	27**	23*
LEVEL OF ASPIRATION: REALISM	50**	06	16	15
MEAN INTERVIEW RATING	16	32**	23*	26**

[a] DECIMAL POINTS OMITTED FROM COLUMN ENTRIES

$^{*}p \leq .05$

$^{**}p \leq .01$

SLIDE #10

SETS OF MAXIMALLY PREDICTIVE VARIABLES[a]

PREDICTOR SET	AV-Op			AV SPECIAL			OVER			REEN		
	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N
1. NAVY CLASSIFICATION TESTS	23	ARI	71	18	ARI	104	29	ARI	105	26	ARI	105
2. BIOGRAPHICAL VARIABLES	42	YRED	71	41	YRED	104	37	YRED	105	33	YRED	105
3. ASSESSMENT CENTER VARIABLES	62	LEVEL OF ASPIRATION: REALISM	71	54	NUMBER OF COMMENTS	104	44	DUAL TASK: NUMBER OF CORRECT SETTINGS	105	38	COORDINATION SPEED & ACCURACY: ACCURACY	105
	70	DUAL TASK: NUMBER OF CORRECT SETTINGS	71	58	MEAN INTERVIEW RATING	104						
	75	INSPECTION/SORT: ACCURACY FOR DEFECTIVE ITEMS	71									

[a] DECIMAL POINTS OMITTED FROM SHRUNKEN Rs

SLIDE #11

TEST-RETEST RELIABILITIES OF SUPERVISORS' MARKS

VARIABLE	SOURCE[a]	T SAMPLE		D SAMPLE	
		r_{xx}	N	r_{xx}	N
TRAIT SCORES					
PROFESSIONAL PERFORMANCE	O	35**	42	—	—
PROFESSIONAL PERFORMANCE	S	48***	64	43**	50
MILITARY BEHAVIOR	O	16	42	—	—
MILITARY BEHAVIOR	S	33**	64	26	50
MILITARY APPEARANCE	O	58***	42	—	—
MILITARY APPEARANCE	S	48***	64	41**	50
ADAPTABILITY	O	45**	42	—	—
ADAPTABILITY	S	37**	64	27	50
GLOBAL MARKS					
AV-Op	O	55***	42	—	—
AV-SPECIAL	S	48***	64	43**	50
OVER	S	32**	64	32*	50
REEN	S	29*	64	24	50

[a] O = OPERATIONAL, S = SPECIAL QUESTIONNAIRE

*$p < .05$

**$p < .01$

***$p < .001$

ATTENUATED ZERO-ORDER VALIDITIES FOR GLOBAL CRITERIA

	CRITERION[a]			
VARIABLE	**AV-Op**	**AV-SPECIAL**	**OVER**	**REEN**
OPERATIONAL VARIABLES				
ARI	35*	30*	53**	52*
ETST	32*	12	18	18
YRED	43**	49**	34*	33
ASSESSMENT CENTER SCORES				
PREDICTED PERFORMANCE IN RECOMMENDED ASSIGNMENT	39*	12	16	15
NUMBER OF DISCUSSIONS AND VOTES	26	30*	28	39*
NUMBER OF COMMENTS	34*	56**	30	32
INSPECTION/SORT: ACCURACY FOR DEFECTIVE ITEMS	-35*	-07	-18	-24
TOOL AND OBJECT NAMING: SCORE 1	20	13	19	39*
DUAL TASK: NUMBER OF CORRECT SETTINGS	44**	38**	41*	22
COORDINATIVE SPEED AND ACCURACY: ACCURACY	24	23	44**	39*
PERCENTAGE ACCURACY	32*	40**	48**	43*
LEVEL OF ASPIRATION: REALISM	67**	09	28	28
MEAN INTERVIEW RATING	22	46**	48**	41*

[a] DECIMAL POINTS WERE OMITTED FROM THE VALIDITY COEFFICIENTS.

*$p \leq .05$
**$p \leq .01$

SLIDE #13

SETS OF MAXIMALLY PREDICTIVE ATTENUATED VARIABLES[a]

PREDICTOR SET	AV-Op			AV-SPECIAL			OVER			REEN		
	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N	SHRUNKEN R	PREDICTOR	N
1. NAVY CLASSIFICATION TESTS	33	ARI	71	28	ARI	104	52	ARI	105	51	ARI	105
2. BIOGRAPHICAL VARIABLES	59	YRED	71	60	YRED	104	67	YRED	105	65	YRED	105
3. ASSESSMENT CENTER VARIABLES	85	LEVEL OF ASPIRATION: REALISM	71	80	NUMBER OF COMMENTS	104	81	DUAL TASK: NUMBER OF CORRECT SETTINGS	105	76	COORDINATIVE SPEED & ACCURACY: ACCURACY	105
	97	DUAL TASK: NUMBER OF CORRECT SETTINGS	71	86	MEAN INTERVIEW RATING	104	88	COORDINATIVE SPEED & ACCURACY: PERCENTAGE ACCURACY	105	83	NUMBER OF DISCUSSIONS & VOTES	105
				89	COORDINATIVE SPEED & ACCURACY: ACCURACY	104				87	MEAN INTERVIEW RATING	105
				94	DUAL TASK: NUMBER OF CORRECT SETINGS	104				92	INSPECTION/SORT: ACCURACY FOR DEFECTIVE ITEMS	105

[a] DECIMAL POINTS OMITTED FROM SHRUNKEN Rs

USING AN ASSESSMENT CENTER TO PREDICT LEADERSHIP COURSE PERFORMANCE OF ARMY OFFICERS AND NCOs

Frederick N. Dyer
Richard E. Hilligoss
Army Research Institute Field Unit
Fort Benning, Georgia

INTRODUCTION

The assessment center concept involves the immersion of an individual in situations which simulate those he would face if he were selected for entry or promotion and assessment of his performance in this simulation. It has been widely used in industry and business to select personnel for high level positions.[1] In 1973-1974 the U.S. Army Infantry School (USAIS) Assessment Center (ACTR) assessed students from the Infantry Officer Advanced Course (IOAC), the Infantry Officer Basic Course (IOBC) and the Advanced NCO Educational System (ANCOES) to determine the feasibility of the assessment center as a technique for leadership development and leadership prediction. It also assessed students from the Branch Immaterial Officer Candidate Course (BIOCC) to determine the feasibility of the assessment center concept as a selection device.[2] Dyer and Hilligoss[3] related the ACTR scores on these Officers and NCOs to ratings of field leadership obtained six months following completion of leadership training and assignment to new duty stations. These ratings were made by supervisors, peers, and subordinates of the former assessee. Prediction of this field leadership criterion was poor. In fact, the more assessor time that went into assessment of the individual the poorer the correlation with this field leadership rating criterion for that exercise. This was true despite high reliabilities of both the ACTR measures[4] and the field leadership ratings. The latter was indicated by high correlations between the 6-month ratings and ratings made on the same individuals at 18 months following assignment to new units. Self-description instruments did a much better job than ACTR exercise assessor ratings in predicting the leadership ratings. It appeared that the ratings made by subordinates, peers, and superiors were strongly influenced by the leader's self-perception of his leadership skills.

---

[1] Earles, J. A. and Winn, W. R. Assessment Centers: An Annotated Bibliography. AFHRL -TR-77-15, May 1977.

[2] U.S. Army Infantry School. Assessment Center After Action Report: Executive Summary (Book 1, Vol. 1), December 1974.

[3] Dyer, F. N. and Hilligoss, R. E. Using an Assessment Center to Predict Field Leadership Performance of Army Officers and NCOs. Proceedings of the 19th Annual Conference of the Military Testing Association, October 1977.

[4] Smith, K. H. Behavioral Assessment of Leadership Skills. U.S. Army Research Institute. December 1975.

The purpose of the present paper is to examine the utility of the ACTR measures for prediction of another criterion, namely, the end-of-course grade obtained by the assessee in the leadership course that he completed immediately after going through the assessment center.

METHOD

ASSESSMENT CENTER PERSONNEL

The assessors consisted of six Majors, seven Captains, two Lieutenants, three Master Sergeants, two Sergeants First Class, and one Staff Sergeant. The assessors were selected by DA using the following criteria: each man must be in one of the combat arms; each Captain and above must have had command experience; each Major, Captain, and Sergeant must have served in combat; and Officers must have an advanced degree in one of the behavioral sciences. The assessors received training for four months on principles and techniques in assessment, interviewing and counseling before beginning their duties. The training included repeated rehearsals of assessment exercises.

Table 1 presents a summary of assessee characteristics and group sizes. Assessees reported to Fort Benning one week before their scheduled USAIS course to participate in the assessment center. They were randomly selected by DA from all students scheduled for USAIS leadership training.

ASSESSMENT CENTER EXERCISES

The ACTR staff, with assistance from Army Reearch Institute and HumRRO scientists, constructed exercises and questionnaires to measure ten dimensions of leader behavior. Leadership research indicated these dimensions to be appropriate for the assigned mission and it was believed these dimensions could be evaluated using the assessment center concept. These were adaptability, administrative skills, communication skills, decision making, forcefulness, mental ability, motivation, effectiveness in an organizational leadership role, social skills, and supervisory skills. In evaluating possible exercises and exercise concepts, a basic factor of consideration was that the exercises would place the assessees in uniquely

Table 1

ASSESSEE GROUP CHARACTERISTICS AND SIZES

	ASSESSMENT GROUP			
Descriptor	IOBC	IOAC	BIOCC(OCS)	ANGOES
Number Assessed	90	88	143	87
Number completing leadership courses	87	84	105	79
Pay Grade	0-1	0-3	E 3-6	E 6-7
Average Age	22.6	28.8	25.3	33.3
Average years of Active Duty	0.3	5.7	3.3	12.9

different situations while simultaneously providing multiple opportunities for the evaluation of each dimension. Exercises were developed which exhibited situational diversity, military relevance and apparent potential for eliciting behaviors related to the designated dimensions.[5] The following exercises were developed:

Entry Interview: A background interview to elicit information related to motivation, experience and the assessee's self-knowledge of his strengths and weaknesses (Time: 65').

Appraisal Interview: An applied exercise in which each assessee interviewed two others to select one for a position within a battalion. This interview elicited behaviors related to communication skills, social interaction and organization of thought (105').

Leaderless Group Discussion: This exercise was a combined individual and group task in which 6 IOAC assessees were assigned a mission to distribute year-end funds among the represented directorates while attempting to acquire a maximum amount for his own directorate. IOBC, BIOCC, and ANCOES assessees were assigned a mission to get a soldier from their unit selected as the Brigade Soldier of the Month and providing a rank order of merit list of the available candidates. This exercise elicited behaviors associated with forcefulness, persuasiveness, organizational ability and group interaction (140').

---

[5] Olmstead, J. A., Cleary, F. K., Lackey, L. L., and Salter, J. A. Development of Leadership Assessment Simulations. Human Resources Research Organiztion TR 73-21, September 1973.

In-Basket Exercise (Three versions: IOAC - assessee was placed in the role of a battalion commander; IOBC/BIOCC - assessee was placed in the role of a company commander; ANCOES - assessee was placed in the role of a 1st Sergeant). An in-basket containing many items typical of the appropriate position was presented to the assessee who had 3 hours to address each item in the in-basket. This exercise elicited behaviors relating to problem solving, decision making, work organization and leadership. It was followed by an interview to discuss reasons for action taken and the relationship perceived to exist among some of the actions (Exercise 180′; Interview 80′).

War Game (IOAC assessees only): This was an assigned-role rotating leader exercise conducted in two 160 minute sessions. Teams of 6 players engaged in cost effectiveness analysis in a military force planning environment. Total costs, R&D, intelligence acquisition, balanced offensive/defensive forces were all considered under limited budget and time constraints. This exercise elicited organizational and leadership behavior (Exercise 320′; Orientation 90′).

Radio Simulate (Three versions: IOAC assessees were placed in company commander role; IOBC/BIOCC assessees were placed in a platoon leader role during a civilian emergency situation to insure that lack of military experience did not preclude them from participation in the exercises; ANCOES assessees were placed in the role of acting platoon leaders). It was a 5-hour exercise using radios as the only means of communication. It elicited organizational and leadership behaviors (Exercise 300′; Orientation 90′).

Assigned Leader Group Exercise (Field Exercise) (IOBC, BIOCC, ANCOES): This was a 5-hour rotating leader designated exercise involving a team of 6 assessees. There were 6 lanes with a different obstacle provided for each lane. It elicited emergent leadership, planning and organizational behaviors (300′).

Management Exercise ("Conglomerate"): This was a two hour exercise divided into two planning and two trading periods. The 18-man assessment group was organized into three 6-man groups who competed against each other. This exercise elicited behaviors relating to emergent leadership, aggressiveness and social interaction (120′).

Writing Exercise: This was an exercise designed to measure accuracy of information provided, grammar, spelling and completeness. The IOAC assessees responded to a Staff Action Paper and the other assessment groups to a discharge action (60′).

828

PSYCHOMETRIC TESTS AND SELF-DESCRIPTION INSTRUMENTS

A survey of tests in general was made revealing many possibilities for adoption into the assessment program. The primary criterion for selecting specific tests was relevance of the variables to be tested to the leadership dimensions of administrative skills, communication skills, supervisory skills, forcefulness, adaptability, decision making, and mental ability.

Additional criteria used in selecting tests were: non-offensive test items, suitability in content and format for use with mature adults, adequacy of normative data and theoretical discussions, recency of publication or revision and efficiency in test administration.

Both cognitive and non-cognitive tests were selected specifically to (1) allow for the comparison of an individual score with normative data and (2) verify the results of other assessment measurements. Group tests were selected in order to minimize the number of assessors and the amount of time required for each assessment. The psychometric tests and self-descriptive instruments selected are listed below. The Person Description Blank was developed for this project. All others are described in the Mental Measurement Yearbook.[6]

1. Leadership Opinion Questionnaire
2. Watson-Glaser Critical Thinking Appraisal
3. Nelson-Denny Reading Test
4. Henmon-Nelson Test of Mental Ability
5. Leadership Q-Sort Test
6. Social Insight Test (Chapin)
7. Work Environment Preference Schedule (Gordon)
8. Strong Vocational Interest Blank
9. Edwards Personal Preference Schedule
10. Person Description Blank

---

[6] Buros, O. K., The Seventh Mental Measurements Yearbook. Gryphon Press, Highland Park, N.J., 1972.

Questionnaires to obtain specific background information about the assessee, and to solicit the assessee's opinion of his assessment experience, were also developed. The purpose of these questionnaires was to assist in the overall research effort and to collect suggestions for improving Assessment Center techniques and administration.

CONDUCT OF THE ASSESSMENT CENTER

Assessment activities occupied three-and-one-half days of the assessee's time. Days typically began at 0700 with activities continuing to 2100. This allowed collection of a great deal of information in the short time available, enhanced the "total immersion" experience, and reduced the effects of outside influences on ACTR performance. Paper and pencil tests, simulated leadership tasks and interviews were approximately equally distributed over the three-and-one-half-day period. Certain groups of assessees returned for feedback counseling from one to three weeks following their assessment. During this three-hour period their leadership strengths and weaknesses, as identified in the assessment center, were communicated and activities were suggested which would lead to correction of deficiencies.

LEADERSHIP COURSE PERFORMANCE

The assessees were all students in USAIS Leadership Courses and attended these courses immediately after their assessment. The courses ranged in length from 12 weeks for the Infantry Officer Basic Course (IOBC) and the Advanced NCO Educational System (ANCOES) through 14 weeks for the Branch Immaterial Officer Candidate Course (BIOCC) to 36 weeks for the Advanced Infantry Officer Course (IOAC). Table 2 lists the number of hours which were devoted to different subjects in each of these courses.

Tables 3 and 4 illustrate the number of examination points associated with different activities. The total possible score was 1000 for each of the courses. Actual means and standard deviations for the total scores obtained by the assessees are given in Table 5. No data are available for the variances of subtests of the total score and it is thus impossible to accurately estimate how much each subtopic added to the total score. However, the points of the subtest probably reflect to some measure its contribution.

For the most part the instruction was conducted by the lecture method and testing was traditional paper and pencil multiple choice. The exceptions are the military stakes and PT testing of the IOBC curriculum.

Table 2

Academic Hours for Four ACTR Groups

Title	IOBC	IOAC	BIOCC	ANCOES
Combined arms subjects	282.5	510.0	100.0	102.0
Staff subjects	27.0	193.0	44.0	119.0
General subjects	83.5	117.5	188.0	106.5
Communications/Electronics	10.0	23.0	11.0	15.0
Unit/Materiel readiness	42.5	44.0	23.0	16.0
Weapons	73.0	44.0	50.0	18.0
Student Evaluation & Counselling	36.0	100.0	105.0	20.0
Electives	-	45.0	-	42.0
Guest Speaker program	-	18.0	-	-
	554.5	1094.5	521.0	438.5

Table 3

Composition of Total Score for IOBC and IOAC Groups

IOBC

Subject	Points
Map reading	10
Pro facts	50
Land navigation (field)	120
Leadership	100
Mil stakes Part I*	140
Mil stakes Part II*	170
Patrolling	10
Patrolling evaluation	100
Army Physical Fitness Test (APFT)	100
Communication/maintenance	100
Written Performance	100
	1000

IOAC

Subject	Points
Medical services support quiz	10
Indoor land navigation	25
Leadership management	45
Staff functions	125
Nuclear, Chemical, Biological operations (NCB)	35
Maintenance management	55
Engineer	10
Communications	25
Fact sheet	10
Disposition Form	10
Cmt 2 to Disposition Form	10
Arty	25
Graphics quiz	10
Operations	30
Company tactical oper, field	80
Company tactical oper, field	75
Company tactics	25
Bn defense	50
Bn offense	50
Internal defense dev	30
Aerial employment	35
Memorandum	10
Staff study	40
Response to nonconcurrence	10
Indorsement military ltr	10
Final Comp Part I	50
Bde defense	30
Bde offense	30
Final Comp Part II	50
	1000

---

*"Hands-on" performance test of various equipment.

Table 4

Composition of Total Score for BIOCC and ANCOES Groups

BIOCC

Subject	Points
Squad drill performance	60
Platoon drill	60
Oral presentation	50
Land navigation field exam	15
Phase I Comp	120
Land navigation field	120
Maintenance management	100
Phase II Comp	175
Army Physical Fitness Test (APFT)	100
Phase III Comp	200
	1000

ANCOES

Subject	Points
Land navigation outdoor	40
Land navigation indoor	40
Communications	40
Graphics	10
Leadership Group, Medical	55
Weapons	95
Maintenance	70
Combat Support	85
Mechanized Training	70
Forward observer	80
Fire direction control (FDC) I	90
Writing Req Mil ltr	15
FDC II	80
FDC III	85
Spot Quiz	10
Fundamentals of Tactics	35
Cmt 2 to Disposition Form	15
Staff	85
	1000

Table 5

Mean and SD for Total Scores

Group	N	Mean	Standard Deviation
IOBC	87	857.84	41.56
IOAC	84	839.74	47.10
BIOCC	105	876.53	46.52
ANCOES	79	810.38	54.41

TABLE 6

CORRELATIONS WITH THE CRITERION

OF ASSESSOR RATINGS FOR THE LEADERLESS GROUP DISCUSSION

LGD Dimensions	IOAC	IOBC	BIOCC	ANCOES
	Assessment Group			
Initial Presentation				
Formal oral communication	.26**	.26**	.23*	.19*
Oral organization	.34**	.24*	.20*	.10
Presentation impact	.24*	.18*	.08	-.02
Group Discussion				
Participation	.05	.09	.06	.20*
Group leadership/ facilitation	.12	.07	.25**	.24*
Persuasiveness	.17	.14	.28**	.24*
Convey information/ communication	.27**	.10	.23**	.31**
Social Concern	.05	.05	.27**	.07

*.05, **.01

805

RESULTS

The scores obtained from the ACTR fall into the following six classes:

1. Assessor ratings of assessee performance during individual and group formal exercises such as the In-Basket,

2. Peer rankings of assessees in those formal exercises where a group of assessees participated together such as the Assigned Leader Group Exercise,

3. Self-rankings by the assessee of his performance relative to other group members in these group exercises,

4. Leadership dimension ratings made by an assessor during the Entry Interview with the assessee,

5. Assessee performance on paper and pencil performance tests, and

6. Assessee self-descriptions on questionnaires and other instruments such as the Edwards Personal Preference Schedule.

The results will be discussed for each of the above classes of score and, following this, the classes of ACTR scores themselves will be discussed and compared on their effectiveness for prediction of the field leadership ratings criterion. Proportions of successful predictors will be compared among classes as will the amount of time required by assessors and assessees to obtain each successful measure. The end result will be an ordering of the different classes of ACTR measure on their utility for predicting the criterion.

1. ASSESSOR RATINGS OF ASSESSEE PERFORMANCE DURING FORMAL EXERCISES

Leaderless Group Discussion (LGD)

The Leaderless Group Discussion rating form was in two parts: Initial Presentation Rating (IPR), three items, and Discussion Participation Rating (DPR), six items. Correlations with the criterion for assessor ratings during this exercise are presented in Table 6. "Formal oral communication" was the only significant dimension common to all four assessment groups. "Oral organization," and "conveys information" each had significant correlations with the criterion for three out of the four assessment groups. Only "negative social impressions," failed to have a significant correlation with the criterion for any of the assessment groups.

Conglomerate Exercise (CONG)

The Congolomerate Game Rating Form consisted of eight items. Each of the assessor ratings had a significant correlation with at least one of the assessment groups. These correlations are presented in Table 7. "Leadership emergence," "energy and vigor," and "decision quality" each had significant correlations with the criterion for three of the four assessment groups. "Receptivity," and "group facilitation," predicted the criterion only for the ANCOES assessees and "sensitivity," only for the BIOCC assessees.

Radio Simulate

The Radio Simulate Leadership Dimension Rating Form is divided into two parts: Platoon (P) and Battalion (B), each having the same eight items. The Platoon rater was an Assessor who acted as subordinate to the assessee in the exercise. The Battalion rater acted as his supervisor. Table 8 presents the correlations with the criterion for assessor ratings on these items. "Decision making," predicted the criterion for all assessee groups in the Platoon ratings. "Communication skills," predicted the criterion for all assessee groups in the Battalion ratings. "Communication skills," and "motivation" predicted the criterion for three of the four assessee groups in the Platoon ratings. "Adaptability," "motivation," "forcefulness," and "administrative skills" predicted the criterion for three of the four assessee groups in the Battalion ratings.

In-Basket

Thirteen of the fourteen dimensions showed significant correlations with the criterion for at least one of the assessment groups. These correlations of assessor ratings with the criterion are presented in Table 9. "Supervision of subordinates," "attention to detail," and "task orientation," were significantly correlated with the criterion for all assessee groups. "Decision making," "use of available information," and "working with superiors," were significantly correlated with the criterion for three of the four assessee groups. "Written communication," was significantly correlated with the criterion for IOBC assessees only, and "self-confidence," for ANCOES only. "Sensitivity," did not correlate significantly with the criterion for any assessee group.

Appraisal Interview

The Appraisal Interview consisted of eight dimensions; five of which predicted the criterion. These correlations are given in Table 10. "Self-confidence," "use of information," and "accomodation," did not predict the criterion for any of the assessee groups.

Writing Exercise

TABLE 7

CORRELATIONS WITH THE CRITERION OF ASSESSOR RATINGS FOR THE CONGLOMERATE EXERCISE

CONGLOMERATE DIMENSIONS	Assessment Group			
	IOAC	IOBC	BIOCC	ANCOES
Energy & Vigor	.25*	.19*	.09	.25*
Leadership Emergence	.22*	.12	.22*	.28**
Oral Communication	.21*	.05	.15	.29**
Decision Quality	.27**	-.04	.18*	.43**
Sensitivity	.15	.13	.22*	.18
Receptivity	.09	-.04	.06	.26*
Group Facilitation	.17	.16	.15	.27**
Overall Effectiveness	.27**	.19	.12	.23*

*.05, **.01

TABLE 8

CORRELATIONS WITH THE CRITERION OF ASSESSOR RATINGS FOR THE RADIO SIMULATE

RADIO SIMULATE DIMENSIONS	IOAC	IOBC	BIOCC	ANCOES
Social Skills P	-.04	.12	.37**	.22*
Social Skills B	-.07	.23*	.26**	.17
Communication Skills P	-.01	.33**	.31**	.35**
Communication Skills B	.21*	.29**	.39**	.30**
Adaptability P	.14	.16	.24**	.42**
Adaptability B	-.00	.28**	.18*	.29**
Motivation P	.13	.18*	.24**	.22*
Motivation B	.04	.20*	.21*	.28**
Forcefulness P	.21*	.17	.29**	.11
Forcefulness B	.11	.21*	.33**	.20*
Decision Making P	.21*	.32**	.39**	.35**
Decision Making B	-.02	.09	.19*	.33**
Administrative Skills P	.04	.08	.23*	.40**
Administrative Skills B	.11	.34**	.18*	.32**
Effectiveness in Org. Leadership Role P	.10	-.00	.35**	.28**
Effectiveness in Org. Leadership Role B	.13	.11	.25**	.29**

*.05, **.01

TABLE 9

CORRELATIONS WITH THE CRITERION OF ASSESSOR RATINGS FOR THE IN-BASKET EXERCISE

IN-BASKET DIMENSIONS	Assessment Group IOAC	 IOBC	 BIOCC	 ANCOES
Written Communication	.11	.21*	.05	.15
Planning & Organization	.31**	.19*	.16	.18
Supervision of Subordinates	.29**	.21*	.19*	.33**
Task Orientation	.30**	.25*	.24**	.22*
Decisiveness	.42**	.22*	.08	.12
Working with Superiors	.21*	.19*	.22*	.13
Personal actions and Initiative	.29**	.20*	.15	.12
Decision making	.37**	.25**	.09	.22*
Attention to Detail	.28**	.26**	.18*	.21*
Problem Analysis	.32**	.12	.11	.25*
Directing Ability	.27**	.16	.17*	.13
Use of Available Information	.31**	.14	.34**	.21*
Self-confidence	.10	.07	.04	.37**

*.05, **.01

TABLE 10

CORRELATIONS WITH THE CRITERION OF

ASSESSOR RATINGS FOR THE APPRAISAL INTERVIEW

APPRAISAL INTERVIEW DIMENSIONS	Assessment Group			
	IOAC	IOBC	BIOCC	ANCOES
Topic Selection	.22*	.20*	.11	.23*
Written Communication	.28**	.21*	.30**	.30**
Written Organization	.26**	.32**	.20*	.33**
Planning	.32**	.12	.25**	.29**
Oral Communication	.08	.14	.06	.25*

*.05, **.01

The Writing Exercise consisted of four dimensions. Three of the dimensions were correlated significantly with the criterion for at least one of the assessment groups.

For the IOBC assessees the criterion scores were related to "accuracy," ($r=.24$, $p<.05$) and "completeness" ($r=.31$, $p<.01$).

For the IOAC assessees, only "grammar" was significant ($r=.21$, $p<.05$).

Three of the four dimensions predicted the criterion for the BIOCC assessees: "accuracy," ($r=.23$, $p<.05$) "grammar," and "completeness," ($r=.30$, $.31$, $p<.01$, respectively).

Only one dimension predicted the criterion for the ANCOES assessees: "completeness," ($r=.28$, $p<.01$).

"Completeness," was the only Writing Exercise dimension predicting the criterion for at least three of the assessment groups. "Spelling," did not predict the criterion for any of the assessment groups.

Assigned Leader Group Exercise (ALGE)

The Assigned Leader Group Exercise rating form was in three parts: Leader Behaviors (four dimensions), Behavior Applicable to both Leader and Follower Roles (three dimensions), and Follower Behaviors (two dimensions).

Three of the Leader Behavior dimensions correlated significantly with the criterion for at least one of the three assessee groups. (The IOAC assessees did not participate in this exercise.) Only one of the Behaviors Applicable to both Leader and Follower dimensions correlated significantly with the criterion for one of the assessee groups. Each of the two dimensions of the Follower Behavior items correlated significantly with the criterion for at least one of the assessee groups.

For the IOBC Lieutenants, the end-of-course criterion was positively related to good assessor ratings on Leader Behavior dimensions, "planning," and "leadership," ($r=.27$, $.21$ $p<.05$, respectively). Of the Follower Behavior dimensions both, "leader emergence," and "group facilitation," correlated significantly with the criterion ($r=.24$, $.24$, $p<.05$, respectively).

For the BIOCC assessees "leadership" and "decisiveness" from the Leader Behavior Group were significantly related to the criterion ($r=.20$, $.18$, $p<.05$). "Physical ability ($r=.18$, $p<.05$) was significant from the Behaviors Applicable to both Leader and Follower. "Group facilitation," was the only significant correlation with the criterion of the "Follower Behavior" dimensions, ($r = .23$, $p < .05$).

For the ANCOES Sergeants, the criterion was positively predicted by two dimensions of Leader Behavior: "planning," and "decisiveness," ($r = .21$, $.20$, $p < .05$).

Of the Leader Behavior Group, "planning," "leadership," and "decisiveness," were significant for at least two of the three assessment groups. "Flexibility," was not significant. Of the Behaviors Applicable to both Leadership and Follower Roles, both "motivation," and "stress tolerance," were not significant. "Physical ability," was significant but for the BIOCC assessees only. Of the Follower Behavior group, "leader emergence," was significant only for the IOBC assessees. "Group facilitation," was significant for two of the three assessee groups.

Leader Game

Only the IOAC Captains participated in this exercise (it took the place of the ALGE for this group). The Leader Game was quite successful in predicting the criterion. In fact, Assessor Ratings for this exercise provided the highest percentage of successful predictors (78%). Of the nine dimensions, seven were successful in predicting the criterion: "organization," "supervisory skills," "participation," "problem comprehension," "leader emergence," and "overall effectiveness," ($r = .31$, $.29$, $.47$, $.33$, $.39$, $.36$, $p < .01$). At the .05 level, "planning," was significant ($r = .25$). "Organization," and "flexibility," were not significant.

2. PEER-RANKINGS ON GROUP EXERCISES

Leaderless Group Discussion

The six group members who participated in this execise ranked all six members on six dimensions at the end of the exercise. Each of the six dimensions was significantly correlated with the criterion for at least one of the four assessee groups. These correlations are presented in Table 11. "Overall effect," was the only variable which correlated significantly with the criterion across all four assessee groups. "Oral communication" "leadership," and "sociability," correlated significantly for at least three of the four assessee groups. "Persuasiveness," correlated significantly with the criterion for the ANCOES assessee group only.

Conglomerate Exercise

Peer ranking correlations with the criterion for this exercise are

TABLE 11

CORRELATIONS WITH THE CRITERION OF PEER RANKINGS FOR THE LEADERLESS GROUP DISCUSSION

	Assessment Group			
DIMENSION	IOAC	IOBC	BIOCC	ANCOES
Oral Communication	$.24^{*}$	$.21^{*}$	$.35^{**}$	.18
Sociability	$.20^{*}$	$.24^{*}$	$.25^{**}$	.12
Leadership	$.23^{*}$	.03	$.18^{*}$	$.31^{**}$
Idea Quality	$.23^{*}$	.06	$.23^{**}$	.11
Persuasiveness	.18	.15	.08	$.29^{**}$
Overall Effect	$.32^{**}$	$.18^{*}$	$.25^{**}$	$.31^{**}$

*.05, **.01

presented in Table 12. No dimension was significantly correlated across all assessee groups. "Popularity," and "acceptance," were significant across three of the four assessee groups. "Conflict," was not significant for any of the assessee groups.

Assigned Leader Group Exercise

This exercise was the least predictive of the four exercises which included peer rankings. However, each of the four dimensions did predict the criterion for at least one of the assessee groups. None of the four dimensions predicted the criterion for the ANCOES. These correlations are presented in Table 13.

Leader Game (IOAC Only)

Peer rankings for the Leader Game were the most predictive peer rankings. In fact all of the five dimensions were significant predictors. These are included in Table 13. These high correlations indicate that assessees ranked highly by peers on the exercise tended to receive the high end-of-course scores.

3. SELF-RANKINGS ON GROUP EXERCISES

Leaderless Group Discussion

The assessee included himself in the group rankings for this exercise and his self-ranking was tested also as a predictor of the criterion. Only four of the six dimensions were found to predict the criterion for the IOAC assessees. These were, "oral communication," "leadership," "idea quality," and "overall effect," ($r = .22$, $.24$, $.23$, $.21$, $p < .05$, respectively).

None of the dimensions predicted the criterion for the other three assessee groups. "Persuasiveness," and "sociability," did not predict the criterion for any of the assessee groups.

Conglomerate

Self-rankings for four of the five dimensions were significantly associated with the criterion on this exercise for the IOAC assessee group. These were "popularity," "planning," "energy," and "acceptance," ($r = .28$, $.43$, $p < .01$, $r = .21$, $p < .05$, respectively).

None of the dimensions predicted the criterion for the other three assessee groups. "Conflict," did not predict the criterion for any of the assessee groups.

TABLE 12

CORRELATIONS WITH THE CRITERION OF PEER RANKINGS FOR THE CONGLOMERATE EXERCISE

DIMENSION	Assessment Group			
	IOAC	IOBC	BIOCC	ANCOES
Popularity	.12	.20*	.23*	.21*
Energy	.34**	.27**	.11	.13
Acceptance	.35**	.20*	.11	.19*
Planning	.28**	.13	.11	.34**

*.05, **.01

810

TABLE 13

CORRELATIONS WITH THE CRITERION OF PEER RANKINGS FOR THE ASSIGNED LEADER GROUP EXERCISE (IOBC, BIOCC, ANCOES) AND LEADER GAME (IOAC)

	Assessment Group			
Dimension	IOAC	IOBC	BIOCC	ANCOES
Social Association	-	.29*	.15	-.01
Leadership	.43**	.23*	.39**	.04
Support of Leader	.39**	.15	.27**	.00
Generating esprit	.24*	.19	.35**	-.17
Problem Comprehension (IOAC only)	.50**	-	-	-
Overall Effectiveness (IOAC only)	.43**	-	-	-

*.05, **.01

Assigned Leader Group Exercise

As for the ALGE Peer Rankings, self-rankings on the ALGE were the poorest predictor of he criterion. Of the four dimensions only "generating esprit," ($r=.21$, $p<.05$) was significant for the IOBC assessees. None of the other assessee groups had any of the dimensions which predicted the criterion. The IOAC assessees did not participate in this exercise.

Leader Game (IOAC Only)

As for the Assessor Ratings and Peer Rankings, he Leader Game Self Rankings were the best self-ranking predictors of the criterion. All of the five dimensions were significant predictors. These were, "problem comprehension," "leadership," "support of leader," "overall effectiveness," and "generating esprit," ($r=.38$, $.36$, $.34$, $.34$, $p<.01$; $r=.19$, $p<.05$, respectively).

4. ENTRY INTERVIEW PERFORMANCE EVALUATION

The correlations of entry interview ratings with the criterion are presented in Table 14. All but two of the fourteen dimensions correlated significantly with the criterion for at least one of the assessee groups. "Goal convergence," and "creativity," did not produce any significant correlations. "Fluency," "asset evaluation," and "liablity evaluation," successfully predicted the criterion for three out of the four assessment groups. "Sense of humor," "task orientation," and "task motivation," correlated significantly for the IOBC assessee group only. "Enthusiasm," and "self-development," correlated significantly for only the BIOCC assessee group.

5. PENCIL AND PAPER PERFORMANCE TESTS

The four test that fall into this category are the Henmon-Nelson Test of Mental Maturity, The Watson-Glaser Critical Thinking Appraisal, the Nelson-Denny Reading Test, and the Social Insight Test.

Since these tests strongly reflect previous academic achievement, it is not surprising that they correlate highly with the end-of-course grade. Correlations of these scores with the criterion are in Table 15. Henmon-Nelson Quantitative, Watson-Glaser Critical Thinking, Nelson-Denny Comprehension, Nelson-Denny Total and Social Insight scores were significant across all assessee groups. Henmon-Nelson Verbal, Henmon-Nelson Total and Nelson-Denny Verbal scores were significant across three of the four assessment groups. Nelson-Denny Reading Rate was significant for the IOAC assessee group only.

845

TABLE 14

CORRELATIONS WITH THE CRITERION FOR ENTRY INTERVIEW RATINGS

Dimension	Assessment Group IOAC	IOBC	BIOCC	ANCOES
Sense of Humor	.10	.26**	.12	-.06
Expression of opinion	.10	.22*	.13	.20*
Task Orientation	-.02	.25*	.06	.03
Asset Evaluation	.32**	.19*	.15	.29**
Liability Evaluation	.23*	.24*	.22*	.13
Task Motivation	.03	.24*	.06	.14
Effectively Conveys Information	.28**	.07	.23**	.16
Fluency	.25**	.11	.27**	.29**
Interest Range	.14	.17	.24**	.21*
Enthusiasm	.02	.11	.20*	.09
Self-Development	.17	.07	.19*	.08
Overall Impression	.12	.26**	.18*	.17

*05, **.01

TABLE 15

CORRELATIONS WITH THE CRITERION FOR PAPER AND PENCIL TESTS

	Assessment Group			
TEST SCORES	IOAC	IOBC	BIOCC	ANCOES
Henmon-Nelson Quantitative	$.62^{**}$	$.19^{*}$	$.33^{**}$	$.47^{**}$
Henmon-Nelson Verbal	$.48^{**}$	-.00	$.29^{**}$	$.40^{**}$
Henmon-Nelson Total Score	$.59^{**}$	.09	$.35^{**}$	$.48^{**}$
Nelson-Denny Verbal	$.44^{**}$	.17	$.27^{**}$	$.41^{**}$
Nelson-Denny Comprehension	$.48^{**}$	$.31^{**}$	$.35^{**}$	$.51^{**}$
Nelson-Denny Total	$.49^{**}$	$.26^{**}$	$.34^{**}$	$.49^{**}$
Nelson-Denny Reading Rate	$.36^{**}$	.10	.04	.16
Watson-Glaser Critical Thinking	$.48^{**}$	$.24^{*}$	$.36^{**}$	$.50^{**}$
Social Insight Test	$.21^{*}$	$.18^{*}$	$.28^{**}$	$.30^{**}$

*.05, **.01

6. SELF-DESCRIPTION INSTRUMENTS

Edwards Personal Preference Schedule (EPPS)

The EPPS did not provide a particularly large number of correlations with the criterion even though each dimension was correlated significantly for at least one of the four assessment groups.

For IOBC assessees, need for "achievement," correlated positively with the criterion ($r=.23$, $p<.05$).

For the IOAC assessees need for "achievement," and "dominance," correlated with the criterion ($r=.18$, $p<.05$; $r=.34$, $p<.01$, respectively). Need for "abasement," showed an inverse relationship between that dimension and the criterion ($r=-.24$, $p<.05$).

For the BIOCC assessee groups, needs for "order," and "succorance" (to have others provide help when in trouble, to seek encouragement from others, etc.) showed an inverse relationship with the criterion ($r=-.23$, $p<$ 01; $r=-.21$, $p<.05$).

The ANCOES assessee groups had the largest number of significant correlations between EPPS dimensions and the criterion. Needs for "abasement," and "nurturance," (to help friends when they are in trouble, to assist others less fortunate, etc.) showed an inverse relationship with the criterion ($r=-.11$, $p<.05$; $r=-.27$, $p<.01$). "Exhibition" and "endurance," needs correlated positively with the criterion ($r=.20$, $.24$, $p<.05$, respectively).

No single dimension predicted the criterion across all four assessee groups. In fact the only dimensions that were significant predictors across even two assess groups were needs for "achievement" and "abasement".

Work Environment Preference Schedule (WEPS)

High scores on this measure "typify individuals who accept authority, who prefer to have specific rules and guidelines to follow, who prefer impersonalized work relationships, and who seek the security of organizational and in-group identification." Three of the assessee groups showed significant correlations on this measure with the criterion of end-of-course grades. For the IOAC, BIOCC, ANCOES assessee groups inverse correlations were associated with the criterion. This inverse relationship would indicate that those assessees readily accepting authority tended to receive low end-of-course grades ($r=-.28$; $-.29$, $p<.01$; $r=-.24$, $p<.05$, respectively).

This test score for the IOBC assessee group did not correlate significantly with the end-of-course-grade criterion.

Leader Opinion Questionnaire (LOQ)

The LOQ provides two scores: Consideration and Structure. BIOCC assessees scoring high on Consideration on the LOQ were more apt to receive a high score on the criterion ($r=.24$, $p<.01$). "Structure," on the other hand was inversely correlated with the criterion ($r=.34$, $p<.01$). No LOQ scores were significant for the other assessee groups.

Leadership Q Sort (LQS)

None of the seven dimensions of the LQS correlated significantly with the criterion for the IOBC assessee group.

For the IOAC assessee gropu, "leadership values," "technical information," and "decision making," correlated significantly with the criterion ($r=.27$, $p<.01$; $r=.19$, $.23$, $p<.05$, respectively). Inverse correlations were obtained for "consideration of others," and "mental health," ($r=-.26$; $-.31$, $p<.01$, respectively). This would indicate that IOAC assessees higher in consideration for others and mental health would tend to have low scores on the criterion.

"Leadership values," and "personal integrity" were LQS dimensions that correlated significantly with the criterion for the BIOCC assessee group ($r=.25$, $p<.01$; $r=.17$, $p<.05$, respectively).

Two of the seven dimensions of the LQS correlated significantly with the criterion for the ANCOES assessment group. These were, "leadership values," and "technical information" ($r=.22$, $.22$, $p<.05$).

None of the LQS dimensions were significant over all four assessee groups. "Leadership values," was significant for three of the four assessee groups. "Personal integrity" and "decision making," were each positively correlated with the criterion for one assessee group, as were "consideration for others" and "mental health", the latter two producing negative correlations with the criterion. The dimension, "teaching and communication," did not correlate with the criterion for any of the assessee groups.

Person Description Blank

Fifty pairs of adjectives were presented to each assessee (e.g., WARY: 1 2 3 4 5 6 7: GULLIBLE) with instructions to rate himself by circling the number that best described his position between these polar adjectives.

Twenty-five of these fifty pairs produced significant correlations with the criterion for at least one of the assessee groups. The pairs of adjectives and their correlations with the criterion for each assessee group are presented in Table 16. Positive correlations indicate that persons who rated themselves higher than average on the rightmost adjective were more apt to receive high end-of-course grades. Negative correlations indicate that persons who rated themselves higher than average on the leftmost adjective were more apt to receive high grades. A negative correlation does not necessarily mean that people were closer to the "1" end of the scale than to the "7" end of the scale. It only indicates that persons who were on the "1" side of the overall average for the item were more apt to be rated high on the criterion.

COMPARISON OF DIFFERENT CLASSES OF ACTR SCORES

Table 17 presents summary data for all assessee groups for the six classes of ACTR scores. It can be seen that the number of scores per assessee (Column 1) varied from 9 for the Pencil and Paper Performance Tests to 75 for the Self-Description Instruments. The assessor time per score (Column 4) showed a very wide variation from 10.9 minutes per score for Assessor Ratings on Formal Exercises to less than one minute per score for the Self-Description Instruments. The latter small time per score reflects the assessor time savings that resulted from presenting the Self-Description Instruments in a group (six assessees) setting. The zero "assessor times per score" that appear for Peer Rankings and Self Rankings reflect the fact that these scores were provided by the assessees and did not require any additional time of assessors beyond that required for the assessor ratings on these exercises. The "assessee time per score" (Column 6) is prorated over Assessor Ratings, Peer Rankings and Self Rankings. Thus only a single figure is shown for this column for these three categories. It can be seen that assessee time per score is relatively long for the Formal ACTR Exercises. Assessee time per score is longest for the Pencil and Paper Performance Tests and shortest for the Self-Description Instruments.

A successful predictor is defined in this report as one which has a correlation with the criterion that is significant at the .05 level. In Column 2 of Table 3 the average number of successful predictors per assessee is given and Column 3 shows the percentage that this is of the total number of scores for the assessee. The assessor ratings of formal exercises represent the most typical ACTR data and their collection is the raison d'etre of an assessment center. The high percentage of predictions from these rating scores compared to interviews and to questionnaires supports the assessment center concept.

Perhaps the most interesting data is in Column 5 where the assessor

Table 16

PERSON DESCRIPTION BLANK (PDB) "YOURSELF" SCORE

CORRELATIONS WITH CRITERION

PDB Descriptor	Assessment Group IOAC	IOBC	BIOCC	ANCOES
Persuasive (1) Unpersuasive (7)	-.24(.014)*	-.09	.03	.12
Noncompetitive (1) Competitive (7)	.14	.20(.032)*	-.01	.14
Clumsy (1) Graceful (7)	-.21(.027)*	.07	-.06	-.04
Understandable (1) Mysterious (7)	-.04	.05	.05	.20(.039)*
Capable (1) Incapable (7)	-.26(.009)**	-.08	-.01	-.08
Smooth (1) Rough (7)	-.03	-.03	.22(.012)*	.05
Insensitive (1) Sensitive (7)	.10	.00	.16(.048)*	.18
Flexible (1) Rigid (7)	-.02	.05	.18(.032)*	-.09
Plodding (1) Brilliant (7)	.19(.042)*	.10	.07	.02
Tactful (1) Blunt (7)	-.06	-.05	.03	.29(.005)**
Tough (1) (Tender)(7)	-.21(.031)*	.00	.13	.14
Wary (1) Gullible (7)	-.26(.008)**	.02	-.09	-.11

*.05, **.01

Table 16 (cont'd)

PERSON DESCRIPTION BLANK (PDB) "YOURSELF" SCORE

CORRELATIONS WITH CRITERION

	Assessment Group			
PDB Descriptor	IOAC	IOBC	BIOCC	ANCOES
Slow (1) Fast (7)	.30(.003)**	-.02	.00	.09
Unintelligent (1) Intelligent (7)	.32(.002)**	.11	.13	.15
Methodical (1) Creative (7)	-.15	.08	-.17(.042)*	-.27(.007)**
Careful (1) Reckless (7)	-.02	.08	.21(.018)*	.23(.020)*
Funny (1) Sober (7)	.11	-.26(.007)**	.03	.04
Leading (1) Following (7)	-.10	-.22(.019)*	.02	.07
Shortsighted (1) Farsighted (7)	.15	-.23(.018)*	-.08	.06
Mild (1) Forceful (7)	.31(.002)**	-.05	.12	-.14
Ambitious (1) Complacent (7)	-.10	-.22(.022)*	.04	.03
Suspicious (1) Trusting (7)	.27(.007)**	.10	-.01	.23(.023)*
Boring (1) Interesting (7)	-.10	.19(.037)*	-.06	.11
Quiet (1) Talkative (7)	-.02	.23(.015)*	.10	-.15
Colorful (1) Colorless(7)	.03	-.11	.08	.25(.015)*

*.05, **.01

Table 17

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORES - ALL ASSESSEE GROUPS COMBINED (END-OF-COURSE GRADE)

Class of ACTR Score	No. of Scores per Assessee	Average Number Successful Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings Formal Exercises	68	37.00	54.41	14.50	26.64	14.24	28.33
Peer Rankings Formal Exercises	15.25	9.00	59.02	0	0	a	a
Self Rankings Formal Exercises	15.25	3.50	22.95	0	0	a	a
Entry Interview	14	5.50	39.28	4.64	11.82	4.64	11.82
Pencil & Paper Performance Tests	9	7.50	83.33	2.96	3.56	17.78	21.33
Self-Description Instruments	75	13.00	17.33	0.30	1.76	1.83	10.54

[a] Peer and self-rankings included with assessor ratings for these calculations.

time per successful predictor for each class of ACTR score is shown. This ranges from slightly less than 2 minutes per successful predictor for the Self-Description Instruments to 26 1/2 minutes per such predictor for the AssessorRatings of Formal Exercises.

Also of interest are the figures in Column 7 of Table 17. This is the assessee time per successful predictor. Although large differences in assessee time per score (Column 6) exist, when successful prediction is considered, there is not a great difference between the different classes of exercise. (From a cost-effectiveness view, assessee time per predictor is probably less critical than assessor time per predictor. High assessor cost was one of the main reasons for termination of the USAIS ACTR.)

Tables 18, 19, 20, 21 provide the data of Table 17 with a separate breakdown by the different assessee groups. In Column 3 of these tables it can be seen that the percentage of successful predictors for the Assessor Ratings ranges from 45% for IOBC assessees, through 51% for IOAC to 60% for both BIOCC and ANCOES.

Table 22 represents another breakdown of the data in Table 17 by separate exercises.

PREDICTION OF END-OF-COURSE GRADE VS. PREDICTION OF FIELD LEADERSHIP RATINGS

Table 23 presents the percentage of successful predictors of the end-of-course grade (Column 2) as given earlier in Table 17 and compares it to the percentage of predictors of the field leadership criterion used in the earlier validation study of the USAIS ACTR by Dyer and Hilligoss (Column 4). In addition, the percentage of successful predictors of the end-of-course grade is given for the different classes of ACTR score when only the assessees who were included in the earlier validation study were considered (Column 3). This reduced the number (N) for IOAC from 84 to 36, the N for IOBC from 87 to 45, the N for BIOCC from 105 to 40, and the N for ANCOES from 79 to 38.

Although this approximate halving of the N for each assessment group reduced the number of successful predictors of the end-of-course grade, there were still more than three times as many successful assessor rating predictors of the grade than of the field leadership ratings, nearly six times as many successful peer-ranking predictors, twice as many self-ranking predictors and three times as many successful paper-and-pencil test score predoctors. The percentage of successful entry-interview and self-description Instrument predictors was about the same for the two criteria.

Table 18

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: IOBC ASSESSEES (END-OF-COURSE GRADE)

Descriptor	No. of Scores per Assessee	Number of Successful Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings	68	31	45.59	14.52	31.84	14.03	34.38
Peer Rankings	15	8	53.33	0	0	a	a
Self Rankings	15	1	6.67	0	0	a	a
Entry Interview	14	7	50.00	4.64	9.29	4.64	9.29
Pencil & Paper Performance Tests	9	5	55.56	2.96	5.33	17.78	32.00
Self-Description Instruments	75	8	10.67	0.30	2.85	1.83	17.13

[a] Peer and self-rankings included with assessor ratings for these calculations.

Table 19

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: IOAC ASSESSEES (END-OF-COURSE GRADE)

Descriptor	No. of Scores per Assessee	Number of Successful Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings	68	35	51.47	14.43	28.04	14.85	24.34
Peer Rankings	16	13	81.25	0	0	a	a
Self Rankings	16	13	81.25	0	0	a	a
Entry Interview	14	4	28.57	4.64	16.25	4.64	16.25
Pencil & Paper Performance Tests	9	9	100.00	2.96	2.96	17.78	17.78
Self-Description Instruments	75	19	25.33	0.30	1.20	1.83	7.21

[a] Peer and self-rankings included with assessor ratings for these calculations.

Table 20

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: BIOCC ASSESSEES (END-OF-COURSE GRADE)

Descriptor	No. of Scores per Assessee	Number of Successful Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings	68	41	60.29	14.52	24.08	14.03	27.50
Peer Rankings	15	9	60.00	0	0	a	a
Self Rankings	15	0	0	0	0	a	a
Entry Interview	14	7	50.00	4.64	9.29	4.64	9.29
Pencil & Paper Performance Tests	9	8	88.89	2.96	3.33	17.78	20.00
Self-Description Instruments	75	12	16.00	0.30	1.90	1.83	11.42

[a] Peer and self-rankings included with assessor ratings for these calculations.

Table 21

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: ANCOES ASSESSEES (END-OF-COURSE GRADE)

Descriptor	No. of Scores per Assessee	Number of Successful Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessor Time per Successful Predictor (min)
Assessor Ratings	68	41	60.29	14.52	24.08	14.03	29.26
Peer Rankings	15	6	40.00	0	0	a	a
Self Rankings	15	0	0	0	0	a	a
Entry Interview	14	4	28.57	4.64	16.25	4.64	16.25
Pencil & Paper Performance Tests	9	8	88.89	2.96	3.33	17.78	20.00
Self-Description Instruments	75	13	17.33	0.30	1.75	1.83	10.54

[a] Peer and self-rankings included with assessor ratings for these calculations.

Table 22

RESULTS FOR SEPARATE ACTR EXERCISES FOR ALL ASSESSEE GROUPS (END-OF-COURSE GRADE)

Descriptor	No. Scores per Assessee	Avg. No. Success Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings							
Leaderless Grp. Discussion	9	4.50	50.00	7.78	15.56	6.67	14.74
Conglomerate	8	4.00	50.00	7.50	15.00	6.67	16.00
Radio Simulate	16	10.50	65.62	29.06	44.29	24.38	37.14
In-Basket	14	8.25	58.93	7.86	13.33	18.57	31.52
Appraisal Interview	8	3.75	46.88	18.54	39.56	26.25	56.00
Writing Exercises	4	1.75	43.75	8.33	19.05	15.00	34.29
Assigned Leader Group Exercise	9	3.33	37.04	16.67	45.00	17.65	56.25
Leader War Game	9	7.00	77.78	15.93	20.48	21.58	24.12
Peer Ranking							
LGD	6	4.00	66.67	0	0	a	a
Cong	5	2.50	50.00	0	0	a	a
ALGE	4	1.67	41.67	0	0	a	a
Leader War Game	5	5.00	100.00	0	0	a	a
Self-Ranking							
LGD	6	1.00	16.67	0	0	a	a
Cong	5	1.00	20.00	0	0	a	a
ALGE	4	0.33	8.33	0	0	a	a
Leader War Game	5	5.00	100.00	0	0	a	a

Table 22 (cont'd)

RESULTS FOR SEPARATE ACTR EXERCISES FOR ALL ASSESSEE GROUPS (END-OF-COURSE GRADE)

Descriptor	No. Scores per Assessee	Avg. No. Success Predictors	% Successful Predictors	Assessor Time per Assessee Score (min)	Assessor Time per Successful Predictor (min)	Assessee Time per Score (min)	Assessee Time per Successful Predictor (min)
Entry Interview	14	5.50	39.28	4.64	11.82	4.64	11.82
Performance Tests							
Henmon-Nelson	3	2.50	83.33	2.22	2.67	13.33	16.00
Nelson-Denny	4	3.00	75.00	1.67	2.22	10.00	13.33
Watson-Glaser	1.	1.00	100.00	8.33	8.33	50.00	50.00
Social Insight	1.	1.00	100.00	5.00	5.00	30.00	30.00
Self-Description Instruments							
Edwards Personal Preference Schedule	15.	2.50	16.67	0.56	3.33	3.33	20.00
Work Environment Preference Schedule	1.	0.75	75.00	1.67	2.22	10.00	13.33
Leadership Opinion Questionnaire	2.	0.50	25.00	1.67	6.67	10.00	40.00
Leadership Q Sort	7.	2.25	32.14	1.19	3.70	7.14	22.22
Person Description Blank	50.	7.00	14.00	0.02	0.17	0.14	1.00

---

[a] Peer and self-rankings included with assessor ratings for these calculations.

Table 23

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORES - ALL ASSESSEE GROUPS COMBINED END-OF-COURSE GRADE VS. FIELD LEADERSHIP RATINGS

Class of ACTR Score	Avg. No. of Scores per Assessee	% Successful Predictors End-of-Course Grade (all course graduates)	% Successful Predictors End-of-Course Grade (N reduced to graduates with complete leadership ratings)	% Successful Predictors Leadership Ratings
Assessor Ratings Formal Exercises	68	54.41	30.88	8.46
Peer Rankings Formal Exercises	15.25	59.02	37.70	6.56
Self Rankings Formal Exercises	15.25	22.95	24.57	11.48
Entry Interview	14	39.28	12.50	16.07
Pencil & Paper Performance Tests	9	83.33	55.56	16.67
Self-Description Instruments	75	17.33	14.67	14.33

DISCUSSION

Two perspectives exist for discussion of these results. One is in terms of the specific characteristics as measured in the ACTR which predict leadership course performance of the different assessee groups. The other perspective for viewing these results is in terms of the general question of what parts of the ACTR were effective in assessment of leadership.

CHARACTERISTICS OF SPECIFIC ASSESSEE GROUPS

The young lieutenant who, following his Basic Infantry Course received a high end-of-course score, judged himself to be less sober than his collegue who did less well in the course. This young officer was also apt to be rated higher in oral and written communication, writing skills, decision quality, attention to detail, adaptability, administrative skills, sense of humor, energetic support of the team effort, and overall good impression than his peer who received a low course grade. He also was apt to be higher in reading comprehension.

The Captain who was about to enter the Advanced Infantry Course and who later received a high end-of-course grade was apt to be more dominant and to have a lower need for order compared to his colleague who received a low end-of-course grade. He viewed himself as more capable, wary, fast, intelligent, forceful, and trusting. His performance was higher in the In-Basket, paper and pencil tests, and the Leader Game. Among the leadership dimensions on which this ACTR Captain was Higher, were planning abilities, overall effectiveness, analysis of problems, supervision, leadership, and decisiveness. He tended to perform better in an unstructured environment. Both mental health and consideration for others showed an inverse relationship with the criterion for these officers.

The enlisted man about to enter the Officer Candidate Course and who received a high end-of course grade was apt to be rated high on the Leaderless Group Discussion exercise, especially on oral communication. He did well also on the Writing Exercise (grammer and completeness). His best exercise was the Radio Simulate, where he received high ratings on social skills, communication, and forcefulness. As with the IOAC assessees, this BIOCC assessee did well on the paper and pencil tests. The BIOCC assessee who had high end-of-course grade was typically rated high on forcefulness, decision making, and use of information. As with the successful IOAC assessee, the successful BIOCC assessee tended to perform better in an unstructured environment.

The NCO about to enter the Advanced NCO Course, who later received a high end-of-course grade, did well on the Radio Simulate. Dimensions on which he did particularly well were communication skills, adaptability, decision making, administrative skills, and effectiveness in an organizational leader role. The NCO that was high on course grades also did well on the paper and pencil tests. This NCO tended to be indifferent to others, and to lack imagination.

PREDICTIVE VALIDITY OF DIFFERENT CLASSES OF ACTR SCORES

The paper and pencil tests provided the largest proportion of criterion predictors, followed by the Formal Exercises (Peer Rankings and then Assessor Ratings). Self-Description Instruments had the smallest propotion of criterion predictors.

It is not surprising that the paper and pencil tests provided the largest proportion of criterion predictors since an end-of-course academic grade reflects, in part, the student's reading and comprehension skills; factors which weighed heavily in the paper and pencil test scores. What is of considerable interest is that the traditional staples of Assessment Centers, i.e., the assessor ratings on formal exercises, predicted this course grade criterion so well. In the previous validation study, those ratings had had almost no predictive validity for the leadership performance ratings which was the criterion measure specifically designed to validate the ACTR.

One other strong contrast exists between the present "end-of-course-grade" validation study and the previous study using field leadership ratings as a criterion of leadership. ACTR performance often was negatively correlated with field leadership ratings of the NCOs. This meant that poor performance on the ACTR often was related to high field ratings for this group. This applied to many assessor ratings and particularly to paper and pencil test scores. Few such inverse correlations were found in the present study (using leadership course performance as a criterion) for the NCOs or for any other group.

An explanation of the failure of the traditional assessment center exercises to predict the field leadership ratings which was proposed in the earlier validation study was that something other than leadership was being rated by the superiors, subordinates and peers who provided these ratings. The success of self-description instruments in predicting the field leadership ratings suggested that little opportunity had existed in these peace-time field settings for leadership to emerge and, in its absence, the leader's self perception was commnicated to the raters and used as the basis for the leadership ratings. In the present study, self-perception measures did much more poorly than assessor ratings in predicting the leadership course grades. Although performance in a largely academic leadership course may not be the best criterion of leadership, the fact that the ACTR formal exercises did predict this criterion, suggests that it may still be a better criterion than the field leadership ratings. Future validation studies will use actual superior ratings (OERs) that bear directly on promotion as a leadership criterion. Promotion itself will also be used as a criterion for some of these future studies.

867

Research Problem Review

VALIDITY OF ASSOCIATE RATINGS OF PERFORMANCE POTENTIAL BY ARMY AVIATORS

Robert F. Eastman
Marie Leger

U. S. Army Research Institute for the
Behavioral and Social Sciences

October 1978

823 808

VALIDITY OF ASSOCIATE RATINGS OF PERFORMANCE POTENTIAL BY ARMY AVIATORS

BACKGROUND

In response to a TRADOC request, the Fort Rucker Field Unit of the Army Research Institute for the Behavioral and Social Sciences has undertaken research to determine attributes which predict aviators who are potentially outstanding combat performers. The effort consists of the following three interrelated tasks: (1) Development of an attack pilot profile from analysis of proven performance (Eastman, Leger and Shipley, 1977); (2) development of a rating form for assessment of potential attack pilots; and (3) selection and evaluation of AH-1 trainees using the findings of tasks 1 and 2.

Currently, no systematic selection of candidates for AH-1 transition training exists. Many trainees are assigned to transition training because they are due for reassignment. A need exists to provide unit commanders with reliable and valid instruments to select aviators for AH-1 transition training. If unit commanders had more and better information, an improved fit of aviators to training assignments would result. The research reported herein is part of task 2 and was conducted to determine the predictive validity of unit level ratings of AH-1 candidates.

OBJECTIVES

The principal objective of this research is to determine the validity of the AH-1 candidate evaluation form as a predictor of trainee performance in the AH-1 transition course.

It was hypothesized that AH-1 (COBRA) qualified pilots in FORSCOM units would be able to predict, by means of assoicate ratings, the AH-1 training performance of aviators from their units. It has already been shown that COBRA pilots in cavalry and attack units demonstrate high inter-rater reliability when evaluating the potential success aviators in their units for AH-1 transition and gunship pilot duties (Eastman and McMullen, 1976). This study will determine validity of the Attack Pilot Candidate Evaluation Form in predicting the flight and gunnery transition grades of AH-1 students. An additional variable of interest was the relationship between length of rater-ratee acquaintance and magnitude of the ratings (Freeberg, 1969; Lewin and Zwany, 1976)

METHOD

SAMPLE

Ratees: The ratees were 45 FORSCOM aviators, all rotary wing qualified and assigned to AH-1 transition training at Fort Rucker. The ratees were selected from AH-1 class rosters if their unit of origin was one with AH-1 aircraft in the TOE. The units were selected on a worldwide basis and are representative of aviation units with COBRA pilots.

Raters: The raters were AH-1 qualified aviators from the units of the AH-1 students. The number of raters in the sample units varied considerably. Because of the requirements of field duty, not all AH-1 qualified aviators were available to evaluate the students from their units. However, no systematic basis for nonavailability which would influence the results of this study was apparent.

Procedure: The AH-1 transition course lasts 5 weeks, and the classes are begun every two weeks. Beginning in Oct 76 when rosters became available for an incoming class they were examined and students arriving from units which were likely to have an attack pilot element were earmarked. The student's unit was then contacted to confirm that a number of COBRA pilots were available. Next a package of rating forms was sent to a point of contact (POC) such as unit XO or a senior attack pilot. The POC then distributed the rating forms and an envelope to all the available AH-1 pilots and later collected them in sealed envelopes to insure confidentiality. Finally, the set of rating forms was returned to ARI in a mailer provided for that purpose. This procedure was followed for all classes during a 14 month period between October 1976 to December 1977. It was necessary to include this large number of classes because only a minority of AH-1 students met the criteria established. Many of the students who could not be used were turnaround Initial Entry Rotary Wing (IERW) students who had just finished flight school. Another large group came from units with no COBRA pilots.

Rating Scale: The rating form used was designed to have raters discriminate among ratees on a set of desirable characteristics for attack pilots. The characteristics rated were identified during structured interviews of 58 attack pilots with combat experience at Ft Knox, Ft Hood and Ft Rucker. On the evaluation form the rater (the AH-1 qualified pilot) is instructed to consider the set of attack pilot characteristics and to assign the AH-1 student a numerical rank, between 1 and 25, representing standing within a typical group of 25 pilots. The rater was also provided space within which to write a 2 - 3 sentence word picture justifying the numerical rating assigned. Additional information was also recorded on where the rating was conducted and the type and duration of the relationship between the rater and ratee. Detailed instructions were provided, some of which only apply when the rating form is to be used to rate a group of AH-1 candidates (see Appendix A).

RESULTS AND DISCUSSION

The median rank order rating was computed for each student from the set of ratings received from his unit. This measure was used to predict two criteria: (1) AH-1 flight transition grades, and (2) AH-1 gunnery grades. The predictive validity of the median rating was determined by computing a Pearson's r between the predictor and each criterion grade. The results in Table 1 show that the validity coefficient for ratings on flight transition grades, r = .32, was high enough to be useful as well as statistically significant ($p<.01$). By contrast, the lower predictive validity of ratings for the gunnery phase of AH-1 transition is probably not useful as a predictor, $r = .21$. The significant difference between these two validities ($p<.01$) may be attributable to differences in the quality of grading the two phases. During the flight transition phase, performance criteria and IP standardization have been established for grading AH-1 students. During the gunnery stage, grading is not based on specified performance criteria, e.g., accuracy is not graded. Improvements in gunnery grading criteria are needed before this training performance can be adequately predicted.

TABLE 1

CORRELATIONS BETWEEN TRANSITION GRADES, GUNNERY GRADES AND THE MEDIAN RATINGS RECEIVED BY AH-1 STUDENTS (N = 45)

Variables	r	p	r^2
AH-1 transition and median rating	.32	<.01	.10
AH-1 gunnery and median rating	.21	<.05	.04
AH-1 transition and gunnery	.33	<.01	

Although the validities obtained are not very high, the predictive validity of .32 accounts for more than 10% of the variance in transition grades and will be useful in selecting AH-1 students. Moreover, the validities reported are a very conservative underestimate of those which would be obtained with an unrestricted population of AH-1 candidates. Because the ratees had already been selected for AH-1 transition, it is reasonable to expect that the ratings of marginal and average aviators were somewhat inflated. This was supported by positive skewing of the distribution of ratings which suggested the use of the median as a datum. Because these data were obtained by mail, the number of ratees was probably fewer than would be possible than if ratings had been conducted as a unit level operational procedure.

The criteria grades for both the transition and gunnery phases are not very discriminating of training performance because of management and grading policies/practices which preclude failures and encourage giving 85s to graduate aviators in advanced training. Some indication of this is provided by the means and standard deviations of flight transition and gunnery grades shown in Table 2. Considering these factors, the .32 validity obtained for prediction of gunnery grades is an encouraging finding in conjunction with the high reliability demonstrated by aviator associate ratings (Eastman and McMullen). Properly used at the unit level, associate ratings would provide a useful selection tool to unit commander and training officers.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF AH-1 TRANSITION
AND GUNNERY GRADE (N = 45)

Phase of Instruction	X	SD
Flight Transition	84.13	3.04
Gunnery	85.93	1.77

No significant relationship was found (r = .09) between the length of acquaintanceship of the rater and ratee and the magnitude of the ratings given.

A related AH-1 Candidate Selection Study included an open ended section in which the rater gave a verbal picture of the ratees. The verbal content of this section was analyzed for those aviators who scored above average in the AH-1 transition. The comments for those who were rated high (above 8.0), or medium (8-15), and low (16-25) are presented in Table 3.

CONCLUSIONS

The validity (r = .32) of ratings in predicting AH-1 flight transition training grades indicates that ratings of potential transition students by COBRA pilots would provide useful information to unit commanders and training officers in selecting aviators for training. The true validity of ratings is anticipated to be somewhat higher than that obtained in this study, because of limitations imposed by the procedures and available sample.

Highly rated good students were regarded to be aggressive leaders while the low rated poor students lacked aggressiveness and did not desire gunship duties. However, factors such as dependability and team performance emphasized by raters appear to contradict the self reported impulsive/independence of the ACE group. The rater received a questionnaire to rate the student identical to the one shown in Appendix A.

TABLE 3

FREQUENTLY OCCURRING REMARKS MADE BY RATERS OF TWO EXTREME GROUPS OF AH-1 TRANSITION STUDENTS

High Rated Pilots Who Obtained High AH-1 Grades[1]		Low Rated Pilots Who Obtained Low AH-1 Grades[2]	
Characteristics	No. of Times Noted	Characteristics	No. of Times Noted
Dependable	22	Lacks aggressiveness	19
Aggressive	20	Lacks dependability	14
Good team worker	18	Does not desire gunship training	14
Has leadership qualities	16	Lacks self discipline	13
Competent	15	Lacks confidence	11
		Poor team worker	10
		Poor performance as an aviator	8

[1] The high group data is based on 5 pilots evaluated by a total of 46 raters.

[2] The low group data is based on 4 pilots evaluated by a total of 34 raters.

REFERENCES

1. Eastman, R. F. and McMullen, R. L., Reliability of Associate Ratings of Performance Potential by Army Aviators ARI, Research Memorandum 76-2, November 1976.

2. Eastman, R. F., Leger, M. and Shipley, B. D., Analysis of Questionnaire Data to Identify "ACE" Attack Helicopter Pilots.

3. Freeberg, N. E., Relevance of Rater-Ratee Acquaintance in the Validity and Reliability of Ratings. *Journal of Applied Psychology*, 1969, 53, 518-524.

4. Lewin, A. Y. and Zwany, A., Peer Nominations: A Model, Literature Critique and a Paradigm for Research. *Personnel Psychology*, 1976, 29, 423-447.

APPENDIX A

Complete this for ATTACK PILOT CANDIDATE EVALUATION

Complete this form only if you are AH-1G qualified.

Instructions:

1. Evaluate this man in your unit/class in terms of your estimate of his potential ability to become a successful gunship/attack pilot. Determine where you think he would rank in a typical group of 25 pilots (number 1 the highest ranking, 25 the lowest ranking). Consider the ATTACK PILOT CHARACTERISTICS below prior to rating each man. Consider the entire group you are asked to evaluate and the following restrictions before beginning. (a) No more than two individuals may be placed in 1-5 column. (b) no two individuals will be assigned the same rating number. Do not rate yourself.
2. Under REMARKS, write a 2-3 sentence word picture to justify the numerical rating you assigned. State briefly the characteristics (desirable or undesirable) of this man that impressed you most.
3. Your ratings will remain anonymous. The packet you picked up has an ID number only to insure that you followed the restrictions when rating.

EVALUATED INDIVIDUAL'S NAME (Last, first)

DATE DAY / MONTH / YEAR

ATTACK PILOT CHARACTERISTICS

DESIRES GUNSHIP DUTIES	AGGRESSIVENESS	CONFIDENCE
TACTICAL KNOWLEDGE	SELF-DISCIPLINE	TEAMWORK
TIMELINESS OF ACTION	DRIVE	INITIATIVE
MECHANICAL ABILITY	EFFECTIVE MAP USE	DEPENDABILITY

CANDIDATE'S PRESENT LOCATION (Circle one)	IERW	UNIT	TRANSITION TRAINING
RELATIONSHIP TO CANDIDATE (Circle one)	HIS CO	IP	IN SAME UNIT

STANDING WITHIN A 25-MAN GROUP (Circle one)

1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

REMARKS:

HOW LONG HAVE YOU KNOWN THE INDIVIDUAL? ________ YEARS ________ MONTHS

RATER ID # 876

USAAVNC(ARI) Fm 1793, 1 Sep 76, prev ed ob.

PERFORMANCE TEST OBJECTIVITY: COMPARISON OF INTERRATER RELIABILITIES OF THREE OBSERVATION FORMATS

William A. Nugent and Gerald J. Laabs

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

The current methods for evaluating performance in the Navy consist of a variety of techniques that often are not adequately assessed in terms of validity, reliability, or objectivity prior to their use as measures of job performance. While some of these procedures may yield useable information related to job performance, the accuracy of that information may be limited. For example, a portion of the performance evaluations conducted in the Navy are based upon a rater's judgment concerning observed performance. It is typically assumed that when such "hands-on" performance tests are conducted, the tests and resultant data are valid and reliable. However, if any ambiguity exists in terms of the performance steps to be observed and evaluated; reduced agreement among all the various raters is sure to result, which seriously affects the validity and reliability of the evaluation procedure. The reduced agreement that occurs among raters in this type of evaluative situation stems primarily from a lack of test objectivity.

Objectivity in performance testing refers to the consistency with which raters make their judgments. One of the important variables that directly affects test objectivity is that of test format. Without specific guidelines on what steps or processes to observe, a rater is forced to make subjective judgments that are based on personal standards and prejudices. Raters should not be expected to evaluate steps they cannot see, such as those involved in evaluating a mental process, and each step should be clearly stated. When several ongoing processes are observed and evaluated as a single step, or there is ambiguity as to what constitutes a performance step, it becomes difficult to obtain consistent ratings across raters. On the other hand, the more structured a test format is, the more the raters should agree on completion of steps in a problem.

Another important variable that may interact with test objectivity is the expertise of the rater. The degree of experience that raters have with a particular piece of equipment will influence their judgments of how others use it. Within the context of the Navy's Personnel Qualification Standards (PQS) program, for example, job performance evaluations are conducted by senior supervisory personnel, or by job incumbents that have successfully passed the section to be evaluated. Unfortunately, it is assumed that when raters are qualified in this manner, questions concerning the objectivity of the hands-on performance test are not relevant.

8

Purpose

The primary purpose of this study was to determine the amount of interrater agreement and reliability of ratings obtained when three structurally different rating formats were used to evaluate the same behavior. In addition, this study examined the relationship between a rater's ability to accurately evaluate the performance of others as a function of his own skill proficiency within a given task area.

METHOD

Stimulus Materials

The appropriate method for determining the consistency of raters' judgments is to hold constant the behavior to be observed and evaluated. One way behavior can be held constant is by videotaping the test behavior so that any variation in evaluation would be due to rater or format differences and not due to test performance differences. Therefore, a videotape was produced in which Navy employees performed four electrical measurements: negative DC voltage measurement, positive DC voltage measurement, and two resistance measures. These measurements were performed using a Simpson Model 260 volt-ohm-meter (VOM) and a Hydrotronics Test Signal Generator. The latter device was specifically designed to provide electronic signals for a previous research project on the use of test equipment (Laabs, Panell, & Pickering, 1977).

The electrical measurement problems were presented to the rater sample three times in the sequence described above. Each type of electrical measurement was performed correctly on only one of the three showings. On the two remaining presentations, the measurements were associated with errors that varied in magnitude. Thus, stimulus materials consisted of 12 videotaped segments in which the four types of electrical measurements were presented in sequential order, while the correct and incorrect performances were presented randomly.

Of the eight videotaped segments that had errors associated with them, only six gave incorrect meter readings. Ratings of these six tape segments were compared to those of the four segments performed correctly. The criterion for the assignment of a pass or fail judgment for each rating form was made on the basis of the meter readings obtained in these ten problems only. The two remaining videotaped segments contained only minor procedural errors and were not included in the present analysis.

The format of the videotaped segments was standardized. Each segment began with a narration of the electrical measurement problem to be performed. Next, the videotape segment showed the steps the examinee used to solve the measurement problem. The videotaped segments ended with the examinee's statement that the problem had been completed, and the examinee's report of the final VOM reading obtained.

Rater Evaluation Forms

One of three different evaluation forms were used by the participants to rate the videotaped performances of the electrical measurements. These forms consisted of a structured, semi-structured, and unstructured format.

878

The unstructured rating format was modeled after a part of the Personnel Qualifications Standards program. This form required the rater to evaluate overall performance, marking a pass or fail for each measurement problem and recording the errors detected. No structured step-by-step procedures were provided to make the evaluations, nor were any criteria specified for a passing performance on any problem.

The semi-structured rating form is similar to forms the Navy Personnel Research and Development Center has developed and used in the past. The form was adapted from a portion of a performance testing program associated with a self-paced test equipment course that is currently administered at the Submarine Training Center, Charleston, South Carolina. This method required the rater to evaluate the videotaped segments against a number of structured areas of performance, assigning a predetermined weighted value to each area. An area of performance often involved more than one procedural step. When the performance was completed, the rater summed the individual point values assigned to the performance areas to determine whether criterion for passing i.e., 7.5 points out of 10 (or 75% correct) had been met.

The structured rating form was developed specifically for this study. It required the rater to evaluate the videotape segments against a series of procedural steps, each consisting of a single behavior. In addition, this form required that each step be performed in the correct sequence to receive a passing score. The VOM equipment face was reproduced on the form so that the position of control settings, the location of lead connections, and the final meter reading obtained could be easily noted or marked on the response form.

To develop the structured rating format, a preliminary version was presented to 12 Sonar Technician Class "A" School instructors from the Fleet Anti-Submarine Warfare School, San Diego, and they were asked to indicate each step of the procedure that was mandatory to achieve a passing performance for each problem. The final version of the structured rating form consisted only of those steps that 85% of the instructors considered essential for passing. Furthermore, there was general agreement among the instructors on the sequential order for the completion of the steps that were retained for each of the four measurement problems.

Rater Expertise

The second independent variable studied was rater expertise. Expertise level was determined by the score the rater obtained on a VOM proficiency test. This test consisted of the same four types of electrical measurement problems that the raters were asked to evaluate during the videotaped presentations. The VOM proficiency test also used the identical equipments as those used in the production of the videotaped segments.

Two proficiency level categories were established for the rater expertise variable: raters who passed two or more problems out of four were considered to be *high skill proficient*; whereas raters who failed to pass at least two of the four problems were considered to be *low skill proficient*. The structured rating form was used by a member of the research staff to evaluate the proficiency level of the raters.

Procedures

Testing was conducted in an experimental laboratory at the Navy Personnel Research and Development Center, San Diego. One half the rater sample received the VOM proficiency test prior to the videotaped presentations; half after viewing the videotape. In both conditions, raters were tested individually on the VOM proficiency test and each rater was assigned to use one of the three rating forms on a random basis.

Raters conducted their evaluations of the videotaped segments in groups of two or three at individual television monitor carrels so that one rater's judgment would not influence another. Prior to evaluating the videotape segments, raters were given a practice session to become familiar with the composition of the videotaped presentations as well as the rating format they had been assigned. The raters viewed each segment, consisting of a single electrical measurement problem, only once. Following each segment presentation, raters were given a 30-second time period to complete entries to their rating forms. The forms were collected when the raters had completed their evaluation of the final videotaped segment.

As discussed previously, the three rating formats differed from one another with respect to the process by which performance steps were observed and evaluated for each videotaped segment. However, the three forms were comparable in that that they provided raters with a means of judging the overall product (i.e., assigning a passing or failing score for each electrical measurement problem). Consequently, the criterion by which the performance of the raters was measured involved comparison of the rater's dichotomous pass/fail responses for each segment to the predetermined standard for the 10 videotaped presentations that were analyzed.

Sample

A total of 15 instructors and 63 students from the Anti-Submarine Warfare School, San Diego, participated in the study. The students in the study were either designated Sonar Technicians or were undergoing Class "A" School training in that rating.

Of the 78 raters tested; 28, 26, and 24 raters were assigned on a random basis to the structured, semi-structured, and unstructured format, respectively. On the basis of the VOM proficiency test, 16 of the raters who used the structured format were classified as high skill proficient and 12 as low skill proficient. Of the raters who used the semi-structured format, 14 were classified as high skill proficient and 12 as low skill proficient. Finally, 12 of the raters who used the unstructured format were classified as high skill proficient and 12 were classified as low skill proficient.

RESULTS

Proficiency Test/Rating Form Presentation Order

No differences were found in terms of criterion agreement with the videotaped presentations as a function of whether the VOM proficiency test was given before or after viewing the videotape. Significant differences also failed to appear in terms of correct performances on the VOM proficiency test as a function of whether the videotaped presentations were shown before or after the VOM proficiency test. Therefore, for all remaining analyses, raters were collapsed across presentation order.

Interrater Reliability

An estimate of interrater reliability was calculated for each form through application of the dichotomous pass/fail responses to an analysis of variance technique that yields an intra-class correlation (Winer, 1971, p. 283). It was found that raters who used the structured rating form showed the highest interrater reliability with a coefficient of reliability of .996. The reliability coefficients for the semi-structured and unstructured formats were .973 and .808, respectively. These coefficient differences were tested by a chi square analysis (Snedecor & Cochran, 1967, p. 286) and were found to be statistically significant ($\chi^2 = 42.4$, df = 2, $p < .001$). Although the structured rating form had the highest coefficient of interrater reliability, the semi-structured and unstructured forms appear to have acceptable levels of interrater reliability in terms of evaluating overall performance on the videotaped segments.

No significant differences were found in interrater reliability values within each rating form as a function of rater skill proficiency.

Criterion Agreement

Table 1 provides a summary of the mean percent agreement with the predetermined pass/fail criterion across the three rating formats. The table shows that the use of the structured rating format resulted in the highest average percent of criterion agreement, while the use of the semi-structured and unstructured formats resulted in progressively less average agreement.

Table 1

Mean Percent Agreement with the Pass/Fail Criterion

Across Three Formats

	Rating Format		
	Structured	Semi-Structured	Unstructured
M	97.1	80.7	76.7
SD	4.6	12.0	14.3

Individual rater percentage values across all three rating forms were converted to standard scores, and an analysis of variance was performed. The main effect of rating format was found to be statistically significant ($F\ (2,75) = 26.34$, $p < .001$). A Scheffe post hoc analysis of the mean values revealed that the structured rating format differed significantly from the semi-structured and unstructured formats at the $p < .01$ level. An estimate of the overall strength of association between rating format and criterion agreement was also calculated. The estimate showed that 39 percent of the variance in the dependent variable can be accounted for by the independent variable.

No significant differences were found in the amount of criterion agreement as a function of rater skill proficiency.

Observation Errors on Failed Problems

The above findings clearly indicate that product judgments (i.e., assigning pass/fail scores) are best made using the structured format. However, these data do not fully describe the state of affairs in using the different formats because they do not reflect the errors made in observing the processes or the procedural steps in the electrical measurement problems. For example, the assignment of a failing score that was in agreement with the predetermined criterion could be made for the wrong reason. This might involve an error of omission (failure to identify an incorrect procedural step) coupled with an error of commission (identifying a correctly performed procedural step as incorrect). Although the three formats were, by design, not equivalent in terms of the amount of information related to process judgments, it was felt that a more detailed examination of the errors made when observing the six videotapes of incorrect performances would be useful.

Table 2 shows the average percent of errors of omission for the three formats. For the structured and semi-structured formats, this means that

Table 2

Mean Percent of Errors of Omission for Three Formats

	Rating Format		
	Structured	Semi-Structured	Unstructured
M	7.1	20.2	50.5
SD	8.6	13.3	28.2

an incorrect step was marked as correct or that points were not subtracted for the incorrect step, respectively. For the unstructured format, this means that the error was not written down. Unfortunately, there is no way of determining whether the rater did observe the incorrectly performed step but merely neglected to enter the error on the observation sheet. Thus, the percent of errors of omission for this format might be inflated. Nevertheless, there was a much lower percent of errors of omission associated with the structured format, which supports the findings on criterion agreement across rating formats.

The other error that could occur on the six failure trials is that of commission. For the structured and semi-structured formats, this means a correct step was marked as incorrect or that points were subtracted for a correct step, respectively. For the unstructured format, this means that a correct step was written down as incorrect. Again, there is no way of knowing if other correct steps were observed as incorrect but simply not entered on the observation sheet. Table 3 shows the percent of raters at both skill levels, and within each rating format, that committed at least one error of commission. Inspection of the table shows that skill proficiency of the rater does not appear to make a difference unless the structured format is used to observe the performance. Overall, the structured rating format is associated with the lowest percentage of raters committing errors of commission (46.4%), with the semi-structured and unstructured formats showing much higher percentages of raters committing these errors (92.3% and 95.8%, respectively).

Table 3

Percent of Raters Making Errors of Commission Across Three Formats and Two Skill Categories

	Rating Format		
Skill Category	Structured	Semi-Structured	Unstructured
High	18.8	92.9	91.7
Low	83.8	91.7	100.0

CONCLUSIONS

A drop from almost perfect agreement with the overall pass/fail criterion when raters used the structured rating format to about 77 percent when raters used the unstructured format, demonstrates the importance of providing a list of unambiguous step-by-step procedures to be checked-off when observing hands-on performance. This finding is further reinforced by the fewer errors of omission and commission committed by this group.

It is interesting to note that the relatively poorer showing for the semi-structured and unstructured formats in terms of overall criterion agreement, and errors of omission and commision occurred for both the high skill and low skill proficient groups. This means that being an expert in a given performance area does not necessarily guarantee that all steps in a given job task will be correctly observed and evaluated by raters who used these performance evaluation forms.

The listing of unambigious step-by-step procedures also resulted in high interrater reliability or objectivity for the structured rating format. With less structure in the rating format, there was less objectivity in observing and evaluating both passing and failing performances. In addition, the level of rater skill proficiency became more important on the structured rating form when errors of commission were examined. Significantly fewer raters in the high skill proficient group made errors of commission than in the low skill proficient group ($\underline{t} = 3.38$, df = 26, $\underline{p} < .01$).

This finding suggests that high skill proficient raters are more apt to accurately observe and evaluate the process by which the electrical measurements were performed. The failure to achieve significant differences between high and low skill proficient raters with respect to commission errors on the two remaining formats may be attributed to a lack of specificity in the performance steps to be observed and evaluated. Thus, no matter what the format of the observation form to be used, the skill proficiency of a rater should probably not be ignored.

The unstructured and semi-structured formats are presently in use in the Navy to evaluate hands-on job performance. It is clear that if these rating formats are replaced by more structured rating forms; more reliable, valid, and objective measurements of hands-on job performance would result.

REFERENCES

Laabs, G. J., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Maintenance of the Missile Test and Readiness Equipment (MTRE MK 7 MOD 2) (NPRDC Tech. Rep. 77-19). San Diego: Navy Personnel Research and Development Center, March 1977. (AD-A037 546)

Smedecor, G. W., & Cochran, W. G. Statistical methods (6th Ed). Ames: Iowa State University Press, 1967.

Winer, B. J. Statistical principles in experimental design (2nd Ed.). New Yor McGraw-Hill, 1971.

Prediction of Field Artillery Officer Performance

by

Arthur C. F. Gilbert
Raymond O. Waldkoetter
Anthony E. Castelnovo

A Paper Prepared for Presentation at the
20th Annual Conference of the Military Testing Association (MTA)
Oklahoma City, Oklahoma
October 30 - November 3, 1978

Performance and Training Research Laboratory
U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

Prediction of Field Artillery Officer Performance

Arthur C. F. Gilbert
Raymond O. Waldkoetter
Anthony E. Castelnovo

U. S. Army Research Institute for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

The development of measures for the prediction of Army officer performance requires evaluation of the utility of these measures within different samples. Other research (Gilbert, 1976) focused on the validation of certain indices within broadly defined groups. These groups were the Combat Arms branches, Combat Support branches, and the Service Support branches. This research was designed to explore the predictive value of certain of these measures in the Field Artillery as the beginning of a validation of these predictors in each of the Army career branches. Another aspect involved was to explore the possible relationship between major field of college study to performance on the prediction and criterion measures.

The first objective of this research was to compare the performance of Field Artillery officers on certain cognitive and non-cognitive measures with that of officers in the other Army career branches. The second objective was to determine the effectiveness of these measures in predicting officer performance early in their active duty tour. The third objective was to evaluate differences in performance among officers who pursued different fields of study while in college on the prediction and on the criterion measures.

Procedure

Data were obtained on 610 Field Artillery officers who entered on active duty during the 1973 Fiscal Year and who continued on active duty after completion of the Officer Basic Course (OBC). The Officer Evaluation Battery (OEB) was administered to these officers during the Officer Basic Course. The OEB consists of cognitive and non-cognitive measures; the seven subtests are Combat Leadership (Cognitive), Technical-Managerial Leadership (Cognitive), Career Potential (Cognitive), Combat Leadership (Non-Cognitive), Technical-Managerial Leadership (Non-Cognitive), Career Potential (Non-Cognitive), and Career Intent. The description of the items in each of the subtests of the Officers Evaluation Battery is shown in Table 1. Two criterion measures were used. The first criterion of performance used was the final course grades in the Officer Basic Course. Officer Efficiency Report (OER) ratings obtained during the first year of active duty were used as the second criterion.

---

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U. S. Army Research Institute or the Department of the Army.

Table 1

Officer Evaluation Battery (OEB) Subtests and Description of Items

SUBTEST	DESCRIPTION OF ITEMS
Combat Leadership (Cognitive)	Military tactics; practical skills in a variety of areas ranging from out-door activities to mechanical and electronic applications.
Technical-Managerial Leadership (Cognitive)	History, politics; culture; mathematics; physical sciences
Career Potential (Cognitive)	Technological knowledge relevant to military requirements.
Combat Leadership (Non-Cognitive)	Combat leader qualities, occupational interests, sports interest, outdoor interests related to combat leadership
Technical-Managerial Leadership (Non-Cognitive)	Mathematics and physical sciences skills and interest; urban or rural background; scientific interest and ability; decisive leader qualities; and verbal-social leadership
Career Potential	Clerical-administrative interest, versus white collar interest, combat interest
Career Intent	Intention of making the Army a career choice.

The first analysis involved comparing the mean performance of Field Artillery officers with the mean performance of officers in the other career branches on the seven subtests of the Officer Evaluation Battery (OEB). Two analyses of regression were performed using the seven subtests of the OEB as predictors. In one analysis Officer Basic Course grades were the criterion while in the other, the criterion was the Officer Efficiency Report (OER) ratings earned during the first year of active duty. The sample was then divided into five groups on the basis of major study field pursued by the officers while in college. These five groups were Humanities, Business, Engineering, Physical Sciences, and Social Studies. Analysis of variance was used to evaluate the differences among the five groups on each of the prediction and criterion measures.

Results and Discussion

In Table 2, the means of the sample of Field Artillery Officers are shown and the mean of officers in other branches on the seven subtests of the Officer Evaluation Battery. There were not any differences between the means of the two groups on six of the subtests. The mean for the Field Artillery officers was higher than for other officers on the Career Potential (Non-Cognitive) subtest at the .01 level.

The zero order correlations between each of the subtests of the Officer Evaluation Battery and Officer Basic Course final grades are shown in Table 3 as well as the resulting multiple correlation coefficient. The correlations the OEB cognitive scales with this criterion are all significant at the .01 level. Two of the non-cognitive subtests, Technical-Managerial Leadership (Non-Cognitive) and Career Intent also yield correlations with this criterion that are significant at the .01 level. Two non-cognitive subtests, Combat Leadership (non-Cognitive) and Career Potential (non-Cognitive) yielded low and non-significant correlations with Officer Basic Course final grades. All of the seven scales of the OEB yielded a multiple correlation of .44 with the criterion that was significant at the .01 level.

When the zero order correlations between the OEB subtests with the criterion of 1974 Annual Average Officer Efficiency reports, shown also in Table 3, are evaluated only the Technical-Managerial Leadership (Non-Cognitive) subtest yielded a significant correlation with this criterion. The obtained multiple correlation of .14 was significant at the .01 level.

In Table 4, the means of the five different college majors are presented for the seven OEB subtests. Significant differences among the five groups were obtained at the .01 on six of the seven subtests of the OEB. There were not any differences among the groups on the Combat Leadership (Cognitive) subtest.

TABLE 2

COMPARISON OF FIELD ARTILLERY OFFICERS WITH OFFICERS IN OTHER BRANCHES ON THE OFFICER EVALUATION BATTERY (OEB) SUBTESTS

Variable	Mean	
	Field Artillery (N=610)	Other Branches (N=3,947)
Officer Evaluation Battery (OEB)		
Combat Leadership (Cognitive)	105.14	103.37
Technical-Managerial Leadership (Cognitive)	108.36	106.44
Career Potential (Cognitive)	101.85	101.90
Combat Leadership (Non-Cognitive)	108.30	106.61
Technical Managerial Leadership (Non-Cognitive)	101.51	102.57
Career Potential (Non-Cognitive)**	106.87	103.53
Career Intent	114.92	114.53

**Indicates a significant difference on this variable at the .01 level.

TABLE 3

CORRELATIONS OF THE OFFICER EVALUATION BATTERY WITH THE TWO CRITERION MEASURES

	Officer Basic Course Final Grades (N=576)	1974 Annual OER (N=471)
Combat Leadership (Cognitive)	.34**	.07
Technical-Managerial Leadership (Cognitive)	.32**	.01
Career Potential (Cognitive)	.32**	.09
Combat Leadership (Non-Cognitive)	.08	.09
Technical-Managerial Leadership (Non-Cognitive)	.14**	.14**
Career Potential (Non-Cognitive)	.06	- .01
Career Intent	.15**	- .08
Multiple Correlation	.44**	.21**

**Significant at the .01 level.

Table 4

Means for the Five Groups of College Majors

	Humanities (N=20)	Business (N=106)	Engineering (N=30)	Physical Sciences (N=269)	Social Studies (N=156)
Officer Evaluation Battery					
Combat Leadership (Cognitive)	97.15	101.73	106.27	107.11	104.74
Technical-Managerial Leadership (Cognitive)**	100.10	99.77	115.67	114.68	104.67
Career Potential (Cognitive)**	100.80	99.76	114.27	102.52	99.53
Combat Leadership (Non-Cognitive)**	114.25	104.65	118.90	107.21	110.30
Technical-Managerial Leadership (Non-Cognitive)**	103.50	96.60	112.17	102.68	100.63
Career Potential (Non-Cognitive)**	108.15	94.74	113.37	109.99	109.43
Career Intent **	121.00	117.51	117.20	111.40	118.16
Officer Basic Course Final Grades	96.00	102.11	104.90	98.94	100.13
1974 Annual OER Score	101.17	100.05	97.18	101.39	98.87

**A significant difference among groups on this variable at the .01 level.

On the Technical-Managerial (Cognitive) subtest the Engineering and Physical Sciences majors were favored in that order in terms of average performance; those officers who majored in Business had the lowest mean performance on this subtest. Engineering majors were favored on the Career Potential (Cognitive) subtest while those officers who majored in Social Studies had the lowest average performance on this subtest. Those officers who majored in Engineering and in Humanities had higher average performance on the Combat Leadership (Non-Cognitive) subtest. Engineering majors were favored on the Technical-Managerial (Non-Cognitive) subtest and on the Career Potential (Non-Cognitive) subtests. Those officers who majored in Humanities had the highest mean performance on the Career Intent scale. There was not any significant difference among the five groups on the criterion measures (i.e. Officer Basic Course final grades or the Officer Efficiency Report ratings earned during the first year of active duty).

Results of this research indicate that Field Artillery officers are not any different from officers in the other Army career branches on the cognitive and non-cognitive subtests of the Officer Evaluation Battery (OEB) with one exception. Field Artillery officers have higher scores on the Career Potential (Non-Cognitive) subtest of the OEB which is essentially a measure of interest in clerical-administrative, manual versus "white-collar", and combat type of activities.

The Officer Evaluation Battery (OEB) is a substantial predictor of success in the Officer Basic Course for Field Artillery officers. The predictive utility of the Officer Evaluation Battery is less when used in the prediction of Officer Efficiency Report (OER) ratings but is still significant (as indicated by a multiple correlation of .21, significant at the .01 level).

Differences in performance among officers who pursued different fields of college study on the Officer Evaluation Battery subtests, with the exception of the Combat Leadership (Cognitive) subtest, were obtained. Future research will utilize this finding to obtain more accurate estimates of the predictive utility of the instrument.

References

Gilbert, A. C. F. Efficacy of certain measures in predicting Army officer performance. Paper read at the 19th Annual Conference of the Military Testing Association, San Antonio, Texas, October 17-21, 1977. In *Proceeding's of the 19th Annual Conference of the Military Testing Association*. San Antonio, Texas: Air Force Human Resources Laboratory and USAF Military Occupational Research Branch, 1977.

Symposium:

Innovative Test Validation Strategies

Chairman

Marvin H. Trattner

Participants

Brian S. O'Leary

Kenneth Pearlman

Frank L. Schmidt

John E. Hunter

Marvin H. Trattner

804

Construct Validity

Brian S. O'Leary
U. S. Civil Service Commission

Introduction

The Professional and Administrative Career Examination (PACE) is the examination used for the selection of personnel for over 100 Federal professional and administrative occupations requiring a college degree or equivalent. The written test portion of the PACE measures five abilities which are differentially weighted according to the requirements of each occupation to which they are applied. The five abilities measured in the examination were selected based on an analysis of the requirements of the occupations. A construct validation model was used in the development of the written examination.

Construct Validation Model in the Employment Setting

Few organizations have used a construct validation model with employment tests. Some investigators have employed a construct model within a single occupation. For example, Bownas & Heckman (1977) used a construct model in developing a test for selecting firefighters. To my knowledge, CSC is the only organization which has used a construct model across occupational groups.

At one time the construct model was not well accepted. However, the courts now give it equal weight with the other validity models. Moreover, there appears to be a definite change in the professional climate concerning construct validity. In fact, the American Psychological Association in their comments on the proposed testing guidelines state that the construct validity section is one of the most important in the guidelines.

Perhaps the biggest drawback with the construct model is that the necessary operational steps are not well defined. Cronbach and Meehl's (1955) classic construct model with the large nomological nets may be too complex for practical application. A form of Campbell's (1960) trait validity may be more appropriate for the employment setting.

A common trend in almost all discussions of construct validity involves testing of hypotheses concerning the construct(s) in question. Is the construct in question related to measures of behavior in situations where the construct is thought to be an important variable? Procedures for testing such hypotheses can vary greatly from logical analysis, to correlational studies, to controlled experimental studies.

PACE Development

Several practical testing needs tended to dictate the construct model for the development of the PACE. First, a single test was needed so that an applicant could be considered for more than one occupation without taking a large number of tests. Second, it was hypothesized that many Federal occupations require similar abilities even though the actual duties may differ. Third, it was not technically feasible to conduct separate criterion-related validity studies in all the occupations. Thus, a construct model was employed.

The basic design of the research to develop the PACE, in simplified form, was

1. Analyze occupations to determine what duties are performed by journeymen.

2. Analyze the duties to determine what abilities are important for performing the duties.

3. Select test parts which measure these abilities.

4. Develop a system of differentially weighting the test parts according to occupation requirements.

Selection of Occupations and Identification of Duties Performed

The first step in the development of the PACE written test was to identify the occupations for which the test would be used. From the pool of approximately 120 occupations to be covered in the PACE, twenty-seven occupations which had accounted for approximately 70% of the placements in previous years were selected for study.

The Civil Service Commission classification standards for these 27 occupations were then analyzed to determine the duties, or major job components, performed by incumbents working at the journeyman or full performance level within each occupational series. These duties were reviewed and refined by subject matter experts. Six to 20 duties were identified for each occupation.

Selection of Abilities to be Measured

A tentative listing of the knowledges, skills, abilities, and other characteristics (KSAO's) that were judged to be required in these occupations was developed. The KSAO list was based on a review of the classification standards. The list included KSAO's that had been described in psychological literature as underlying successful job performance and KSAO's that experience with Federal testing had shown to be related to successful job performance. Through a review of the literature, six of these abilities were identified as having potential for inclusion in the written test portion of the PACE.

Development of Weighting System

Subject Matter experts (generally supervisors) in each of 27 occupational series rated the duties performed in their series for their importance to successful performance in the occupation and for the relative amount of time that journeymen spend on each duty. A total of 1,241 subject matter experts rated the duties. These persons also rated the abilities for their importance for successful job performance.

Six Civil Service Commission psychologists, experienced in the use of tests for employee selection, rated the importance of each of the six PACE abilities for measuring the performance of each duty for each of the 27 occupations.

For each occupation, the duty importance and time spent ratings obtained from the subject matter experts and the ability importance ratings obtained from the psychologists were used to weight the abilities to be measured by the subtests of the battery. Scores on the PACE subtests were multiplied by the weights, and the sum of the products used to rank order competitors for an occupation.

Seven weighting patterns emerged for all 27 occupations. One ability (long-term memory) was eliminated since the testing literature did not contain any tests suitable for use in a short-term testing session. When this test was eliminated, six weighting patterns emerged for the 27 occupations, two of the weighting patterns covering 23 of the occupations.

Development of the Ability Measures

Literature in the field of psychometrics was reviewed in order to find ways to measure the abilities. The most important sources of suitable tests were the works of French (1951) and French, Ekstrom, and Price (1963). The questions developed for the PACE correspond to the question types contained in these works. The major differences between the French question and the PACE questions lies in the modifications made to develop a selection instrument which could be objectively scored by machine.

Criterion-Related Validity Studies

As soon as the PACE written test was constructed, follow-up research was begun to further develop the empirical base for technical support of the test and of the entire system of relating abilities to job duties. What we are testing with the criterion-related studies is a system of identifying and weighting ability constructs which underlie job performance. The criterion-related validity studies are performed to test out the system. If the criterion-related validity studies demonstrate empirically that abilities do indeed underlie job performance this lends support for the entire system. It is then not necessary to perform criterion-related validity studies in each specific occupation included in the examination.

A series of studies was planned, in which test scores of job incumbents were to be related to the scores of the same incumbents on certain specifically prepared measures of job performance. The basic design of these studies, for each occupation studied can be outlined as follows:

1. Determine what journeymen do on the job - that is, conduct a job analysis.

2. Use the job analysis to develop measures of job performance.

3. Determine the statistical relationship between performance on the test and performance on the job.

Occupations Studied

Social Insurance Claims Examining. The Social Insurance Claims Examining occupation is unique to the Social Security Administration. Employees within this occupation evaluate claims for retirement and health insurance, calculating applicable rates of annuity after the claim is approved and as benefits are increased by change in the Social Security laws. Claims Authorizer is the title for the most complex job type within the occupation. Claims authorizers work only on the initial claim, evaluating its legitimacy and calculating the amount of benefits to be paid.

Internal Revenue Officers. Internal revenue officers investigate delinquent taxpayer accounts, both individual and corporate. The revenue officer must secure and analyze financial information such as profit and loss statements, sales and expense figures, or market value of taxpayer's property.

Revenue officers are empowered to institute levies, attach taxpayers' income, and seize and sell taxpayers' property. Before resorting to such enforced collection action, revenue officers explore alternative methods such as arranging for installment payments.

Customs Inspection. The mission of the Customs Service is to assess and collect customs duties on imported merchandise, to prevent fraud and smuggling, and to control carriers, persons, and articles entering and departing the United States. Customs enforces its own as well as some 400 laws and regulations for 40 other Federal agencies. The primary function of the customs inspector is to process people and merchandise coming into the U. S., to protect the revenue against fraud and theft, and to keep items harmful to our welfare out of the country. Customs inspectors work at airports, seaports, and border points processing passengers and cargo.

Job Analysis

In each occupation a detailed job analysis was conducted through the use of a task inventory, a listing of the tasks performed by job incumbents. Journeymen in each occupation identified the tasks performed in these occupations. Claims authorizers identified 528 tasks, internal revenue officers identified 260 tasks, and customs inspectors identified 494 tasks.

Journeymen were then asked to indicate whether or not they performed each task and to indicate the relative amount of time spent on each. This rating was made on a seven point relative-time-spent scale ranging from "very much below average" to "very much above average."

Responses to the task inventory were analyzed by means of the Comprehensive Occupational Data Analysis Program to determine the relative amount of time spent in performing each task by all journeymen. The relative amount of time spent in performing each task is a measure of its relative importance. An additional analysis was performed in the customs inspector and claims authorizer samples to determine if all journeymen in the sample were performing similar tasks.

Measures of Job Performance

Results from the task inventory were used in the development of the measures of job performance for each occupation. Four measures of job performance were developed.

Job Information Test. In each study the job information test was a multiple choice test requiring one hour to complete. Items for the tests were developed by subject matter experts in the field and were designed to measure the job knowledge required to perform the duties on which the journeymen spend the greatest amount of time.

Work Samples. Work samples are designed to be relevant approximations to the work actually performed on the job. In the claims examiner study the work sample consisted of a standardized claim which had to be adjucated. The claims examiner was instructed to treat the claim as one that he would receive during the performance of his regular duties and to take the necessary appropriate action that he would normally take.

The work sample in the internal revenue officer study consisted of five taxpayer delinquent accounts in which the revenue officer had to make various collection decisions (e.g., seize property, levy). The case folders contained sufficient information to make the necessary collection decisions and closely resembled the actual case folders used in the Internal Revenue Service.

For the customs inspector study a novel videotype simulation was developed. Four sequences of customs activities were shown (e.g., passenger processing, vessel clearance, search, seizure, and arrest). Upon completion of each sequence the customs inspectors were required to complete appropriate customs documents, identify mistakes made during the televised sequence, and recommend proper performance.

Each work sample required one hour and fifteen minutes to complete.

Supervisory Rating Form. The supervisory rating form was a tailor-made rating form designed to record a first-level supervisor's rating of the performance of the subordinate journeymen. The rating scales were developed to correspond to the duties identified in the task analysis. Each supervisor rated his journeymen on different categories of performance for each of the major duties identified in the task inventory. Scale points describing effective and ineffective performance were developed for each scale on the rating form.

Supervisory Ranking Form. The supervisory ranking form contained the same description of the job duties as the supervisory rating form but contained no scale points describing effective and ineffective performance. Each supervisor had to rank his subordinates with respect to each the major duties identified for each occupation. This criterion measure was not used in the internal revenue officer study.

Success in Training. Training success measures were available for a sample of claims examiners. Training success was measured by averaging five training performance measures administered during the five phases of the training program. These training performance measures included actual work samples (i.e., working on actual disability claim) in addition to the traditional multiple-choice type questions.

Research Participants

Two hundred and thirty one claims authorizers, 305 internal revenue officers, and 190 customs inspectors at various locations throughout the U. S. were administered the PACE and the criterion instruments. The total testing time for each participant was approximately 8 hours.

Relationship Between PACE and Job Performance

The total score on the PACE test was significantly related to job performance as measured by all the measures of job performance for the claims authorizer and internal revenue officer studies. For the customs inspector occupation, PACE scores were significantly related to performance on the job information test and the work sample but not the supervisory ratings and rankings. The pattern of validity coefficients was similar across occupations with a median coefficient of .40. These results indicate that persons who score high on the PACE tend to perform better on the job.

Comparisons were also made of different procedures for weighting the subtests of the PACE. The construct weights which are being used operationally produced validities that were essentially as high as those obtained by other weighting procedures.

The correlation obtained between PACE and training success for claims examiners indicates that PACE is a valid predictor of training success.

These highly consistent results provide further support for the construct validity of the weighting system used in the development of the PACE.

Test of a New Model of Validity Generalization:
Results for Tests Used in Clerical Selection

Kenneth Pearlman and Frank L. Schmidt
U.S. Civil Service Commission
John E. Hunter
Michigan State University

The purpose of our research program on validity generalization has been to test one of the orthodox doctrines of personnel psychology: the belief in the situational specificity of employment test validities (Schmidt & Hunter, 1977). This belief has been founded on the empirical fact that considerable variability is observed from study to study in raw validity coefficients even when jobs and tests appear to be similar or essentially identical (Ghiselli, 1966). The explanation that has developed for this variability is that the factor structure of job performance is different from job to job and that the human observer or job analyst is simply too poor an information receiver and processor to detect these subtle but important differences. Until recently, most industrial psychologists accepted this explanation and concluded that empirical validation is required in each situation, and that validity generalization is essentially impossible (Albright, Glennon, & Smith, 1963, p. 18; Ghiselli, 1966, p. 28; Guion, 1965, p. 126). Our work has tested the hypothesis that the outcomes of validity studies within job-test combinations is due to statistical artifacts. This presentation first describes the validity generalization model used to test this hypothesis and then describes the model's application to clerical tests and jobs.

Figure 1 shows how various statistical artifacts might act to produce the appearance of wide variability in validities when in fact none really exists. This figure shows what the observed variability in validity coefficients across studies would be if in fact the true score correlation between test and criterion were equal at .60 in each setting and all variability in results from study to study were due solely to various statistical artifacts.

The first distribution in Row 1 shows the variability to be expected if only the artifact of differences between studies in criterion reliability were operating. The distribution of criterion reliabilities assumed is shown in Table 1.

The second distribution shows variability due solely to differences between studies in test reliability. The distribution of test reliabilities assumed is shown in Table 2.

The third distribution in Row 1 shows variability due solely to differences between studies in degree of range restriction. Range restriction values used in the computations are shown in Table 3.

The single distribution in Row 2 shows the variability produced by the three artifacts in Row 1 operating simultaneously. Even though we have not yet introduced sampling error, it is obvious that observed variability from study to study is already substantial. The distributions in Row 3 show how artifactual variance increases still further when ordinary sampling error is added. The three distributions illustrate expected variability when studies are all based on sample sizes of 50, 100, and 150, respectively. The distributions based on N's of 50 and 100 are probably the most realistic. These standard deviations are very similar to empirically observed standard deviations, as we will see in this study.

Figure 1 illustrates the effects of only four artifactual sources of variance:

1. differences between studies in criterion reliability;
2. differences between studies in test reliability;
3. differences between studies in range restriction; and
4. sampling error (i.e., variance due to $N < \infty$).

There are at least three additional artifactual sources of variance:

5. differences between studies in amount and kind of criterion contamination and deficiency;
6. computational and typographical errors; and
7. slight differences in factor structure between tests of a given type (e.g., arithmetic reasoning tests).

The full variance-components model resulting when all of the above sources of artifactual variance are considered is outlined in Appendix A.

How could one test the hypothesis of situational specificity with real data? Conceptually, this test is quite simple. Suppose, for example, a researcher had 100 validity coefficients relating tests of perceptual speed to proficiency in clerical work. He or she need only convert the validities to Fisher's z, compute the variance of this distribution, and subtract variance due to each of the artifactual sources from this total variance. If one finds that artifacts account for all or essentially all of the variance, the hypothesis of situational specificity is rejected. If this is the case, validity generalization is obviously no longer a problem, since the observed variation in validity results will have been shown to be a result of the operation of statistical artifacts.

Method

Compilation of Validity Distributions

The process of compiling a data base of sufficient scope and size to permit a large-scale test of the model was undertaken in two stages: first, we developed a classification and coding system that would enable us to capture all potentially relevant data from validity studies; second, we made an extensive search of published and unpublished validity studies and recorded the information in these studies according to our coding system. We selected clerical occupations as one of our initial areas of investigation because of the large number of validity studies that have been conducted on such occupations.

Tests were classified using a system partially adapted from Ghiselli (1966, pp. 15-21) and Dunnette (Note 1). This system is shown in Appendix B. Ten general categories of test types were established, most of which represent a construct or ability factor found in the psychometric literature (e.g., verbal ability, quantitative ability, perceptual speed). Categories for general intelligence tests (consisting of verbal, quantitative, and abstract reasoning or spatial ability components), so-called "clerical aptitude" tests (consisting of verbal, quantitative, and perceptual speed components), performance tests (e.g., typing or dictation tests), and motor ability tests (consisting of various types of finger, manual, and arm dexterity tests), were included because of their relatively common use in clerical selection, even though they do not represent pure constructs in the factor analytic sense. Within each general test type category codes were developed for the specific item types most commonly used as measures of that factor or test type (e.g., the verbal ability test type category included such item type categories as reading comprehension, vocabulary, grammar, spelling, and sentence completion).

Clerical jobs were classified using a slightly modified version of the Dictionary of Occupational Titles (DOT) classification system (U.S. Department of Labor, 1965; Pearlman, Note 2). This coding scheme is shown in Appendix C. Under this system clerical jobs are grouped into five "true" job family categories (DOT occupational divisions 20, 21, 22, and 23, plus job groups 240-243 of occupational division 24), one "miscellaneous" category (DOT job group 249), and two additional categories developed to handle clerical occupations which were not sufficiently specified in the original study to permit definitive classification, and samples representing two or more different job families.

905

We collected data only from studies which met certain minimum requirements, including the reporting of: 1) validity results in the form of a bivariate correlation coefficient uncorrected for either attenuation or range restriction; 2) sufficient information to classify the test and job studied; 3) sample size; and 4) sufficient information to classify the criterion as a measure of either job proficiency or training success. Data from studies using such administrative criteria as turnover, absenteeism, and tardiness were not included.

The data collection process included an extensive search for both published and unpublished validity studies of clerical jobs. In addition to a thorough search of the published literature, we reviewed most of the major commercial test manuals for validity information, utilized computer search services, called and wrote test publishers to obtain unpublished validity data, and contacted research groups, private consulting firms, individual psychologists, and government and military personnel psychologists. We ultimately succeeded in locating 3,300 validity coefficients for a variety of clerical jobs and tests. These represented 669 independent samples. Approximately two-thirds came from unpublished studies. Of the 3,300 coefficients, 2,718 are based on overall job proficiency or performance criteria and 582 are based on criteria of training success. Analysis of the validities based on training criteria is not included in this study.

Data Analysis

The validity data were keypunched, entered into a computer file, and sorted into frequency distributions according to the job and test type categories into which they had been classified. The distribution of validity coefficients across the eight job categories and ten test types is shown in Appendix D. Within the five categories of "true" job families, 33 validity distributions were sufficiently large to permit analysis.

To compute the mean and variance of each of our empirical validity distributions, each coefficient was converted to Fisher's z form and weighted by its associated sample size to produce more accurate estimates of these two parameters. The correction for variance due to sample size was thus a weighted average of the sampling error across studies, i.e.,

$$\Sigma[N_i \;/\; (N_i - 3)] \;/\; \Sigma N_i .$$

The information necessary to determine actual values of criterion reliability, test reliability, and range restriction is not presented in the vast majority of research studies (Jones, 1950). Thus one must rely on reasonable assumed distributions of these effects. The distributions of criterion reliabilities, test reliabilities, and range restriction effects assumed in this study are those shown in Tables 1, 2, and 3, respectively. These distributions are probably somewhat conservative (Schmidt & Hunter, 1977), leading to underestimates of variance due to these three statistical artifacts. Criterion reliabilities are for job performance or proficiency measures, not measures of success in training. The model used in the present study is an improvement over the model used in Schmidt and Hunter (1977); unlike the earlier model, the present model includes a correction for variance due to between-study differences in test reliability.

The procedure by which we computed estimates of variance due to between-study differences in criterion reliability, test reliability, and range restriction effects for each validity distribution are presented in Appendix A. After computation, all four estimates of artifactual variance (the above three sources plus variance due to sampling error) were subtracted from the observed variance, providing the final estimate of true situational variance, i.e., variance due to true differences between jobs in the factor structure of performance.

No corrections have been made in our research for differences between studies in amount and kind of criterion contamination or deficiency, for computational and typographical errors, or for slight differences between tests in factor structure because it is difficult if not impossible to estimate their effects. However, not correcting for these sources of error insures a conservative procedure, i.e., the corrected variance tends to overestimate rather than underestimate true variance.

Results and Discussion

Table 4 compares the empirically observed standard deviations of the 33 validity distributions with the standard deviations predicted solely on the basis of test and criterion unreliability effects, range restriction effects, and sampling error. Also shown is the percent of observed variance in each distribution accounted for by these four artifacts, and the total sample size and number of validity coefficients on which each distribution is based.

In 10 of the 33 cases, the predicted standard deviations are slightly larger than the observed standard deviations. These are exactly the type of results we would expect if the situational specificity hypothesis is false. Within a given set of validity distributions representing a variety of job family-test type combinations there are likely to be some distributions in which the three unassessed sources of variance are present to varying degrees and others in which these sources are negligible. In distributions of the former type we would expect the predicted standard deviation to fall below the observed standard deviation to varying degrees. In distributions of the latter type the predicted standard deviation would be expected to fall slightly below the observed standard deviation about half the time and to slightly exceed the observed standard deviation about half the time as a result of minor differences between the actual artifactual effects and our estimates of them.

Considering these distributions together, in only five of the 33 cases is the percentage of observed variance accounted for less than half, and in only one case is it less than 40 percent. The average amount of variance accounted for is 75 percent. This means that, in general, the variance left within which situational specificity (situational moderators) can operate is extremely limited. For many of the distributions, no variance is left. In 20 of the 33 distributions, more than 70 percent of the observed variance is accounted for.

If we look only at the true constructs--eliminating motor ability tests, performance tests, general intelligence tests, and clerical aptitude tests--the average amount of variance accounted for is 84 percent. If we could correct for all seven artifactual sources of variance--instead of just four--we conclude all observed variance would be accounted for.

Thus the evidence is strong that the doctrine of situational specificity is false and employment test validities can be generalized across settings.

Although not shown in Table 4, our method also produces estimates of the true validities that should be generalized. These are produced by correcting the mean observed validity for range restriction and criterion unreliability using average values of both. For the true constructs in Table 4, these validities range, with one exception, from .37 to .70. The average value is .47. Thus tests of these kinds have generalizable and substantial validity for predicting proficiency in clerical work.

We believe that application of this model may lead to fairly dramatic progress in the establishment of general principles and theories about trait-performance relationships in the world of work. The first step in the development of general principles and theories in this or any other area is the establishment of stable patterns of relationships among basic variables. In order to establish such patterns of relationships, it is first necessary to demonstrate that the doctrine of situational specificity is false or essentially false.

If the situational specificity hypothesis is rejected, then it follows that various constructs--for example, verbal ability--have invariant population relationships with specified kinds of performances and job behaviors. The best estimate of this population value for any construct-performance combination is the fully corrected mean of the validity distribution. This mean should be corrected for unreliability in both test and criterion, since the goal in theoretical research is to reveal relationships among underlying constructs, independent of measurement problems. We predict that such research will reveal that the underlying structure of reality in personnel psychology--that is, the pattern of population parameters and their relationships--is considerably simpler than has previously been imagined (Schmidt & Hunter, 1978). The model presented here thus provides a tool which should enable the field to move beyond a mere technology to the status of a science.

Reference Notes

1. Dunnette, M. D. Validity study results for jobs relevant to the petroleum refining industry. Washington, D.C.: American Petroleum Institute, 1972.

2. Pearlman, K. Clerical test validity review project: An interim status report. U.S. Civil Service Commission, Personnel Research and Development Center, January 1977.

910

References

Albright, L. E., Glennon, J. R., & Smith, W. J. *The use of psychological tests in industry*. Cleveland: Howard Allen, 1963.

Brogden, H. E., & Taylor, E. K. A theory and classification of criterion bias. *Educational and Psychological Measurement*, 1950, *10*, 159-186.

Ghiselli, E. E. *The validity of occupational aptitude tests*. New York: Wiley, 1966.

Guion, R. M. *Personnel testing*. New York: McGraw-Hill, 1965.

Jones, M. H. The adequacy of employee selection reports. *Journal of Applied Psychology*, 1950, *34*, 219-224.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529-540.

Schmidt, F. L., & Hunter, J. E. Moderator research and the law of small numbers. *Personnel Psychology*, 1978, *31*, 215-232.

Thorndike, R. L. *Personnel selection*. New York: Wiley, 1949.

U.S. Department of Labor. *Dictionary of occupational titles* (3rd ed.). Washington, D.C.: U.S. Government Printing Office, 1965.

Wolins, L. Responsibility for raw data. *American Psychologist*, 1962, *17*, 657-658.

Figure 1

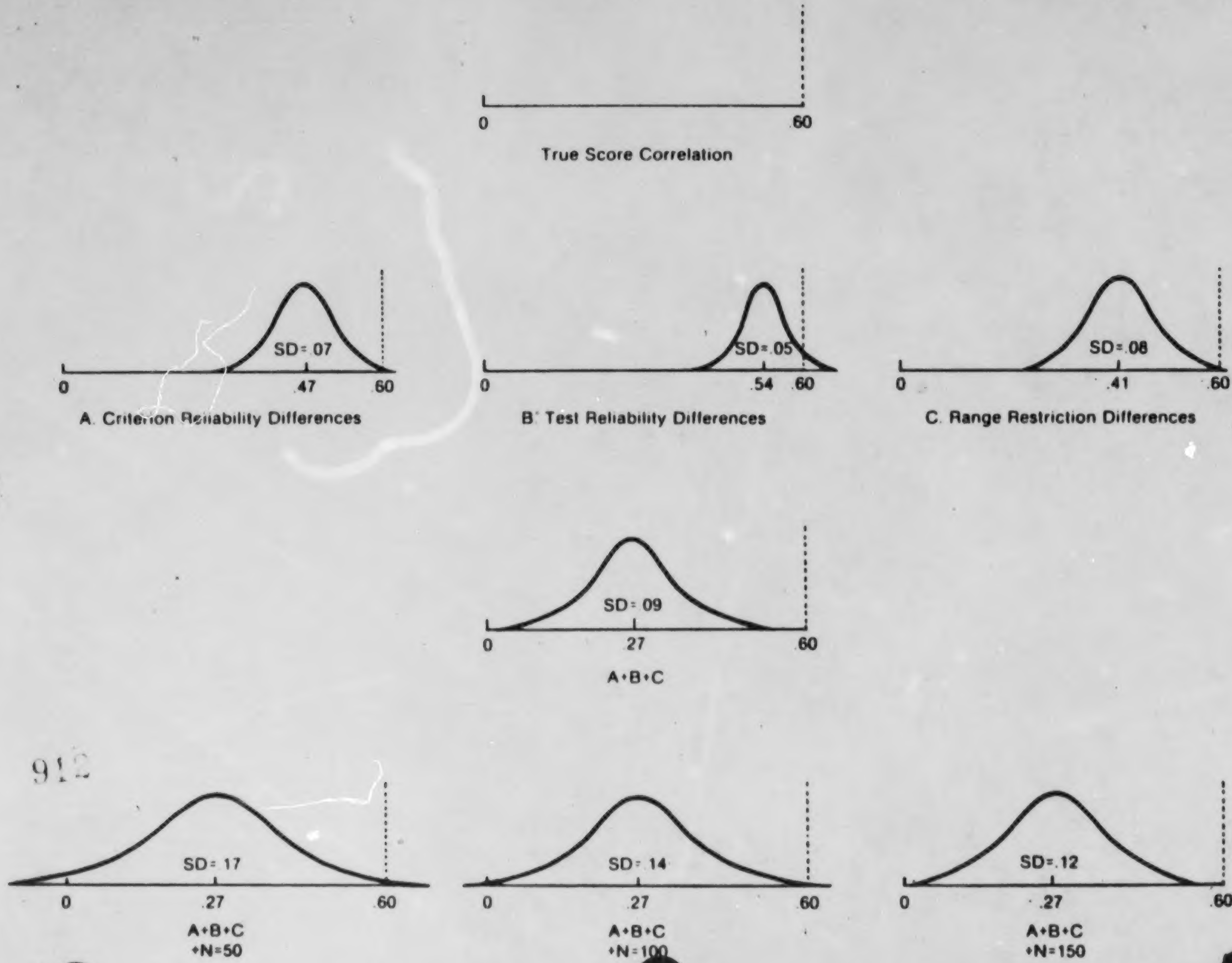

0
.60
True Score Correlation
SD=.07
0
.47
.60
A. Criterion Reliability Differences
SD=.05
0
.54
.60
B. Test Reliability Differences
SD=.08
0
.41
.60
C. Range Restriction Differences
SD=.09
0
.27
.60
A+B+C
SD=.17
0
.27
.60
A+B+C
+N=50
SD=.14
0
.27
.60
A+B+C
+N=100
SD=.12
0
.27
.60
A+B+C
+N=150

Table 1

Example of Assumed Distribution of Criterion Reliabilities Across Studies

(Proficiency Measures)

Reliability	Relative Frequency
.90	3
.85	4
.80	6
.75	8
.70	10
.65	12
.60	14
.55	12
.50	10
.45	8
.40	6
.35	4
.30	3

Note. Expected value (criterion reliability) = .60.

Table 2

Example of Assumed Distribution of
Test Reliabilities Across Studies

Reliability	Relative Frequency
.90	15
.85	30
.80	25
.75	20
.70	4
.60	4
.50	2

Note. Expected value (test reliability) = .80.

Table 3

Example of Assumed Distribution of Range Restriction Effects Across Studies

Prior Selection Ratio	SD of Test	Relative Frequency
1.00	10.00	5
.70	7.01	11
.60	6.49	16
.50	6.03	18
.40	5.59	18
.30	5.15	16
.20	4.68	11
.10	4.11	5

Note. Expected value (*SD*) = 6.0.

Table 4

Observed and Predicted Standard Deviations and Percent Variance Accounted For (Clerical Job Families-Proficiency Criteria)

Job Family	Test Type	Total N	No. of r's	Obs. SD[a]	Pred. SD[a]	% Var. Acc. For
Steno, Typing, & Filing	General Intelligence	3,986	65	.266	.174	43
Computing & Account Recording	General Intelligence	5,433	58	.181	.135	56
Steno, Typing, & Filing	Verbal Ability	16,176	175	.179	.130	53
Computing & Account Recording	Verbal Ability	8,670	110	.180	.132	53
Material & Production Recording	Verbal Ability	1,926	45	.155	.178	100
Information & Message Distribution	Verbal Ability	1,073	14	.165	.147	80
Steno, Typing, & Filing	Quantitative Ability	12,368	130	.148	.138	87
Computing & Account Recording	Quantitative Ability	10,631	140	.171	.149	76
Material & Production Recording	Quantitative Ability	1,641	39	.195	.201	100
Information & Message Distribution	Quantitative Ability	1,110	15	.136	.143	100
Public Contact	Quantitative Ability	993	13	.064	.144	100
Steno, Typing, & Filing	Perceptual Speed	23,045	269	.190	.139	54
Computing & Account Recording	Perceptual Speed	22,978	321	.168	.151	81
Material & Production Recording	Perceptual Speed	3,574	64	.145	.163	100
Information & Message Distribution	Perceptual Speed	2,002	27	.168	.156	87
Public Contact	Perceptual Speed	1,151	16	.126	.137	100
Steno, Typing, & Filing	Reasoning Ability	3,497	36	.134	.123	84
Computing & Account Recording	Reasoning Ability	1,556	27	.205	.169	68
Material & Production Recording	Reasoning Ability	1,114	22	.181	.168	86
Steno, Typing, & Filing	Memory	2,471	36	.169	.147	76
Computing & Account Recording	Memory	1,817	33	.154	.156	100
Material & Production Recording	Memory	1,086	22	.154	.175	100
Steno, Typing, & Filing	Spatial/Mech'l. Abil.	2,604	21	.112	.097	76
Computing & Account Recording	Spatial/Mech'l. Abil.	5,265	57	.150	.121	65
Material & Production Recording	Spatial/Mech'l. Abil.	811	18	.160	.184	100
Steno, Typing, & Filing	Motor Ability[b]	4,045	54	.172	.129	56
Computing & Account Recording	Motor Ability[b]	11,948	131	.132	.117	78
Material & Production Recording	Motor Ability[b]	1,968	27	.131	.133	100
Information & Message Distribution	Motor Ability[b]	1,370	19	.219	.147	45
Steno, Typing, & Filing	Performance Tests	3,665	39	.348	.164	22
Computing & Account Recording	Performance Tests	1,427	15	.178	.122	47
Steno, Typing, & Filing	Clerical Apt. Tests[c]	3,915	53	.235	.165	49
Computing & Account Recording	Clerical Apt. Tests[c]	1,645	25	.217	.161	55

---

[a] In Fisher's z form.

[b] Dotting, tapping, etc. tests; also some manual and arm dexterity tests.

[c] Tests comprised of verbal, quantitative, and perceptual speed components.

Appendix A

Sources of Variance in Distributions of Validity Coefficients for a Given Test Type - Job Combination

1. $\sigma_{e_1}^2$ = Error variance due to differences between studies in criterion reliability.

2. $\sigma_{e_2}^2$ = Error variance due to differences between studies in test reliability.

3. $\sigma_{e_3}^2$ = Error variance due to differences between studies in degree of range restriction.

4. $\sigma_{e_4}^2$ = Error variance due to sampling error, i.e., variance due to use of $N < \infty$.

5. $\sigma_{e_5}^2$ = Error variance due to differences between studies in amount and kind of criterion contamination and deficiency (Brogden and Taylor, 1950).

6. $\sigma_{e_6}^2$ = Error variance due to computational, typographical, etc., errors (Wolins, 1962).

7. $\sigma_{e_7}^2$ = Error variance due to slight differences in factor structure of tests measuring the same construct.

8. σ_{ss}^2 = Variance due to true differences in factor structure between criterion measures, i.e., variance due to true situational specificity.

Appendix A (cont'd.)

Our hypothesis is: $\sigma_{\delta\delta}^2 = 0$. An alternative statement of this hypothesis is:

$$\sigma_{\text{total}}^2 - \sigma_{e_1}^2 - \sigma_{e_2}^2 - \sigma_{e_3}^2 - \sigma_{e_4}^2 - \sigma_{e_5}^2 - \sigma_{e_6}^2 - \sigma_{e_7}^2 = 0$$

I. Computing variance due to differences between studies in criterion reliability.

1. Compute mean of the raw validity distribution in Fisher's z (F_z) form and convert to r.

2. Correct this raw r for test and criterion unreliability and range restriction using average values across studies for these three variables. (In this study, average assumed criterion reliability was .60, average assumed test reliability was .80, and average assumed range restriction was to a *SD* of 6.0 from an unrestricted *SD* of 10.0; see Tables 1, 2, and 3 in text.) This provides an estimate of the fully corrected validity $r_{\infty\infty}$.

3. For each value of assumed criterion reliability, r_{cc_i}; compute $r_{\infty\infty} \sqrt{r_{cc_i}}$ and convert this attenuated r to F_z.

4. Compute $\sum Fz_i \cdot n_i$ and $\sum Fz_i^2 \cdot n_i$, where n_i = the relative frequencies of the criterion reliabilities.

5. Variance in *Fz* distribution due to criterion reliability differences of validities is then:

$$\sigma^2_{r_{cc}} = \frac{\sum Fz_i^2 \cdot n_i}{\sum n_i} - \left[\frac{\sum Fz_i \cdot n_i}{\sum n_i}\right]^2$$

II. Computing variance due to differences between studies in test reliability.

1. Compute mean of the raw validity distribution in *Fz* form and convert to *r*.

2. Correct this raw *r* for range restriction and for attenuation due to test unreliability (using average values of both) but not for attenuation due to criterion unreliability. Let this resulting coefficient be symbolized r_∞.

3. For each value of assumed test reliability, r_{xx_i}, compute $r_\infty \sqrt{r_{xx_i}}$ and convert this attenuated *r* to *Fz*.

4. Compute $\sum Fz_i \cdot n_i$ and $\sum Fz_i^2 \cdot n_i$, where n_i = the relative frequencies of the criterion reliabilities.

5. Variance in *Fz* distribution due to differences in test reliability is then:

$$\sigma^2_{r_{xx}} = \frac{\sum Fz_i^2 \cdot n_i}{\sum n_i} - \left[\frac{\sum Fz_i \cdot n_i}{\sum n_i}\right]^2$$

III. Computing variance due to range restriction differences between studies:

1. Compute mean of the validity distribution in *Fz* form and convert to *r*. Correct this raw *r* for mean range restriction but not for attenuation due to either source of unreliability.

2. For each value of the restricted standard deviation, use the following formula to compute the expected restricted *r*:

$$r_i = \frac{u_i R}{\sqrt{u_i^2 R^2 - R^2 + 1}} ,$$

where:

r_i = the restricted validity

R = the unrestricted validity

u_i = sd_i / SD

SD = the standard deviation of the test in the unrestricted group

sd_i = the standard deviation of the test in the restricted group

This formula is obtained by solving Thorndike's (1949, p. 173) Case II formula for r_i. (Thorndike's Case II is the model throughout these analyses; use of Case III would generally produce very similar results.)

3. Convert r_i to *Fz* and compute $\sum Fz_i \cdot n_i$ and $\sum Fz_i^2 \cdot n_i$.

4. Variance due to range restriction differences between studies is then:

$$\sigma^2_{rr} = \frac{Fz_i^2 \cdot n_i}{\sum n_i} - \left[\frac{Fz_i \cdot n_i}{\sum n_i} \right]^2$$

Appendix B

Test Classification System and Code

General Mental Ability (10)

10 = intelligence/adaptability

Verbal Ability (11-17)

11 = verbal ability, nfc[1]
12 = reading comprehension
13 = vocabulary
14 = grammar
15 = spelling
16 = word fluency
17 = sentence completion

Quantitative Ability (20-28)

20 = quantitative ability, nfc
21 = computation (mixed operations)
22 = arithmetic word problems
23 = error location
24 = computation (addition)
25 = computation (subtraction)
26 = computation (multiplication)
27 = computation (division)
28 = graph and table reading

Reasoning Ability (30-36)

30 = reasoning ability, nfc
31 = verbal reasoning (analogies, inference)
32 = abstract reasoning (figure analogies)
33 = logical order of events
34 = letter series
35 = number series
36 = judgment

Perceptual Speed (40-49)

40 = perceptual speed, nfc
41 = name comparison/checking
42 = number comparison/checking
43 = figure comparison
44 = cancellation
45 = filing (numbers)
46 = name and number comparison/checking
47 = coding
48 = alphabetizing or name filing
49 = substitution (letter-digit or digit-symbol)

Memory (50-56)

50 = memory, nfc
51 = memory of oral instructions
52 = classification
53 = coding
54 = substitution (letter-digit or digit-symbol)
55 = number writing
56 = immediate memory

Spatial and Mechanical Ability (60-65)

60 = spatial or mechanical ability, nfc
61 = mechanical knowledge
62 = spatial relations
63 = location
64 = mechanical principles
65 = pursuit

Motor Ability (70-78)

70 = motor ability, nfc
71 = finger dexterity
72 = hand dexterity
73 = arm desterity
74 = tracing
75 = tapping
76 = dotting
77 = mark making
78 = aiming

Performance Tests (80-83)

80 = performance tests, nfc
81 = typing test
82 = dictation test
83 = work sample

Clerical Aptitude Tests (90)

90 = clerical aptitude (combined verbal, numerical, and clerical speed)

---

[1] Not further classifiable or combination of item types within same test type

Appendix C

Job Classification System and Code
(full D.O.T. code in parentheses)

Stenography, Typing, Filing, and Related Occupations

201 = Secretaries (201.368)
202 = Stenographers (202.388)
203 = Typists (203.588)
204 = Correspondence clerks (204.288)
205 = Personnel clerks (205.368)
206 = File clerks (206.388)
207 = Duplicating-machine operators (207.782)
208 = Miscellaneous office machine operators (208.138, 208.588, 208.782, and 208.885)
209 = Stenography, typing, filing, and related occupations, n.e.c.[1] and mixed samples[2]
260 = Clerk (includes office clerk, general clerk, junior clerk, entry- and intermediate-level clerk) (209.588)
261 = Clerk-typist (209.388)
262 = Index clerk (209.588)
263 = Combined samples of clerks, typists, stenographers, and secretaries
264 = Copy holder (209.588) and/or proofreader (209.688)
265 = Pricing clerk (209.588)
266 = Checker II (209.688)

Computing and Account-Recording Occupations

210 = Bookkeepers (210.388)
211 = Cashiers (211.368 and 211.468)
212 = Tellers (212.368)
213 = Automatic data-processing equipment operators (213.382, 213.582, 213.588, 213.782, and 213.885)
214 = Billing-machine operators (214.488)
215 = Bookkeeping-machine operators (215.388)
216 = Computing-machine operators (216.488)
217 = Account-recording-machine operators (217.388)
219 = Computing and account-recording occupations, n.e.c. and mixed samples
270 = General office clerk (includes senior clerk and administrative clerk) (219.388)
271 = Ward clerk (219.388)
272 = Hand transcriber (219.588)
273 = Toll-bill clerk (includes invoice typist) (219.388)
274 = Budget/fiscal clerk (219.388)
275 = Actuarial clerk (insurance) (219.388)
276 = Accounting clerk (219.488)
277 = Coding clerk (219.388)
278 = Combined samples of computing and account-recording machine operators
279 = Combined samples of bookkeeping, accounting, fiscal, and auditing clerks

---

[1] Not elsewhere classified
[2] Samples which represent two or more different job codes from the same job family

Appendix C (cont'd.)

Material and Production Recording Occupations

221 = Production clerks (221.168 and 221.388)
222 = Shipping and receiving clerks (222.138, 222.387, 222.587, and 222.687)
223 = Stock clerks and related occupations (223.387)
224 = Weighers (224.487)
229 = Material and production recording occupations, n.e.c. and mixed samples

Information and Message Distribution Occupations

230 = Messengers, errand boys, and office boys and girls (230.368, 230.868, and 230.878)
231 = Mail clerks (231.588)
232 = Post office clerks (232.368)
233 = Mail carriers (233.388)
234 = Mail-preparing- and mail-handling-machine operators (234.582 and 234.885)
235 = Telephone operators (235.862)
236 = Telegraph operators (236.588)
237 = Receptionists and information clerks (237.368)
239 = Information and message distribution occupations, n.e.c. and mixed samples

Public Contact Occupations

240 = Collectors (240.368)
241 = Adjusters (241.168 and 241.368)
242 = Hotel clerks (242.368)
43 = Direct service clerks (243.368)

Miscellaneous Clerical Occupations

280 = Enumerator/survey worker (249 268)
281 = Library assistant (249.368)
282 = Order clerk (249.368)
283 = Telehone ad-taker (249.368)
284 = Securities clerk (249.368)
285 = Engineering clerk (249.388)
286 = Service representative (includes contract clerk) (249.368)
287 = Claims examiner (249.268)

Additional Categories

250 = All other clerical occupations not otherwise classifiable or not specified
251 = Samples which represent two or more different job codes from different job families

Appendix D

Number of Validity Coefficients in Clerical Validity Data File by Test Type and Job Family (Proficiency Criteria)

	Job Family[1]								
Test Type	20	21	22	23	24	28	25	26	Total
General Intelligence	65	58	9	6	6	14	28	4	190
Verbal Ability	175	110	45	14	4	19	60	8	435
Quantitative Ability	130	140	39	15	13	21	76	11	445
Reasoning Ability	36	27	22	0	3	6	21	0	115
Perceptual Speed	269	321	64	27	16	35	108	18	858
Memory	36	33	22	2	3	7	4	2	109
Spatial/Mechanical Ability	21	57	18	6	3	5	0	1	111
Motor Ability	54	131	27	19	11	13	6	4	265
Performance Tests	39	15	0	0	0	0	1	2	57
Clerical Aptitude Tests	53	25	9	3	0	3	37	3	133
Total	725	856	159	89	59	123	341	53	2,718

---

[1] Job family codes defined:

20 = Stenography, Typing, Filing, and Related Clerical
21 = Computing and Account-Recording Clerical
22 = Material and Production-Recording Clerical
23 = Information and Message Distribution Clerical
24 = Public Contact Clerical
28 = Miscellaneous Clerical (D.O.T. Group 249 jobs)
25 = Unspecified Clerical
26 = Mixed Samples

Synthetic Validity

Marvin H. Trattner

History

Guion (1965) defines synthetic validity as the inference of validity from the predetermined validities of a test for specific components of a job. Guion's approach allows one to infer test validity for an occupation when no test or criterion data are collected for the specific occupation. The approach to be described here enlarges upon Guion's definition by permitting under certain conditions the calculation of the validity coefficient when no test and criterion data have been collected for an occupation. With the use of this approach test validities can be calculated for occupations where it is infeasible for a variety of reasons to conduct traditional studies. The approach to be described is an application of Ernest Primoff's J-coefficient. It is also based on Vern Urry's recent extensions of the J-coefficient formula.

Description

The following are the steps in applying the synthetic validity paradigm.

1. Select the class of occupations for which the test will be used. For the class, select the most populous occupations. The class should consist of occupations in which similar tasks are performed at approximately the same difficulty level. For instance for the clerical class select Clerk Typist, Secretary, File Clerk, Receptionist, Typist, etc.

2. Define the major job duties for the occupational class. A duty is defined as a major segment or component or module of work performed in an occupation. It could be the only work performed in a specific subtype of the occupation. The same duty may occur in several of the different occupations in the class. The following are good examples of clerical duties: take dictation, compose routine correspondence, type simple material, type technical material. The job duty is conceptually similar to the "work behaviors" defined in the new Uniform Guidelines on Employee Selection Procedures.

3. Determine the test validity for measuring duty performance for several occupations in the class. Correlate the test score with duty performance measures separately for the most populous occupations in the class.

4. Calculate the test's synthetic validity coefficient for a specific occupation. The synthetic validity coefficient is the correlation

of the weighted sum of the duty performance scores with the test score after the duty scores are weighted for importance for the specific occupation. Another way to precisely calculate the test's synthetic validity is to weight the individual test by duty validity coefficients for duty importance and sum the weighted validity coefficients with the use of the the correlation of weighted sums formula. The formula gives the correlation of the test score with the sum of the weighted duty performance scores. Once stable test by duty validity coefficients and duty intercorrelation coefficients are obtained for the occupations in a class they can be used to estimate the test's validity for any occupation in the class. The only additional data required to obtain the validity estimate are ratings of duty importance for success in the occupation. The first level supervisors are employed to rate the duties for importance for occupational success.

If all occupational duties are defined and precise estimates of test validities for the duties are obtained then the synthetic validity coefficient is precisely equivalent to the actual test validity coefficient. *It is assumed that the test correlation with duty performance is constant across occupations*. If all major occupational duties are not defined then the synthetic validity coefficient is a lower bound for the actual validity coefficient. If the duty performance scores correlate inconsistently with each other and with the test across occupations in the class then the synthetic validity coefficient cannot be estimated.

Method

In order for the research to succeed, two major problems will need to be overcome. It will be necessary to define a comprehensive set of duties that describes the work performed in the occupational class. Where the same duty is performed in different occupations it should be performed at the same level of difficulty and consist of very similar tasks. The other difficulty is that the validity coefficients for the test for measuring the duties must be consistent and somewhat significant across occupations.

The method to be described should achieve the desired results.

1. *Assemble subject matter experts (SMEs) to define the duties for the class*. The SMEs whould be senior journeymen and first level supervisors in the occupations. First ask the SMEs to define the duties in their own occupation. Then ask the assembled SMEs to generalize duties across occupations. A generalized duty should define the same work tasks at the same level of difficulty occurring in different occupations. The subject matter specific to an occupation should be omitted if it is unrelated to the duty difficulty level. For instance, the subject matter of the technical material that is typed would probably be irrelevant in determining an aptitude test's validity for measuring skill in typing technical material. Consequently reference to the technical material

should be omitted from the duty definition. Where the same duty occurs at different levels of difficulty then the duty should be split into several which describe the differing difficulty levels.

2. *Determine the test validity for measuring duty performance for the class of occupations*.

It will be necessary to correlate test scores with duty performance scores for incumbents in the populous occupations in the class. Since there may be as many as fifty defined duties, most of which would not apply to any one occupation, the only feasible way to measure duty performance would appear to be with the use of a *rating* of duty performance. It would be prohibitively expensive to construct work samples for fifty duties. Furthermore, it would be necessary that the work samples have subject matter content that would be equally familiar to all research participants. Work samples with neutral subject matter content *in many cases* would closely resemble the aptitude test for which they were designed as criteria. These kinds of work samples might not be scientifically or legally defensible.

We are all aware that ratings are a very questionable kind of performance measure. When employed as criteria they are less likely to be significantly correlated with selection instruments than other kinds of job performance measures. They are used here not for the sake of convenience but out of necessity. The following are some of the steps we will take to maximize the probability of success for the project.

1. Employ a large N for each occupation.

2. Select occupations for study with very specific performance standards. These would tend to be production oriented occupations.

3. Use research participants at grade levels below the journeyman.

4. Identify research participants only by a code number. In this way we hope to encourage more candid and hence more valid ratings.

5. Obtain performance ratings from the first level supervisors and the research participants themselves. Combine the two sets of ratings to obtain the performance measure. The assumption to be tested is that research participants are best qualified to evaluate their relative performance on the duties and the first level supervisors are best qualified to evaluate the research participants' overall performance level.

6. Carefully scale rating forms.

7. Use impossible end points to eliminate raters who use them.

8. Train raters and involve them in the construction of the rating forms.

9. Get reliability estimates for the ratings by comparing ratings given by present and former first level supervisors of the research participants.

The synthetic validity paradigm can be applied to test selection with multiple regression. The validity coefficient for each test can be synthetically calculated and employed along with the test intercorrelations to select and weight tests in a battery.

If a consistent matrix of significant test validity coefficients for duties can be developed for a class of occupations it is probable that the matrix would be applicable across agencies. It is probably true that variance due to duty performance in different occupations ought to be much greater than variance due to employer. A test that correlates with specific duty performance for one employer should correlate the same way for another. It follows that private industry employers, state and local governments, and Federal agencies could each profitable pool their research and development efforts in constructing synthetically validated test batteries.

REFERENCES

Guion, Robert M. Personnel Testing, New York: McGraw-Hill, 1965.

Primoff, Ernest S. Empirical validations of th[illegible]coefficient. Personnel Psychology, 1959, 12, 413-418.

SECTION 10

STATISTICAL AND MEASUREMENT METHODOLOGIES

930

A Primer of Item Response Theory

(an overview of a book by the same title*)

Thomas A. Warm
U.S. Coast Guard Institute
Oklahoma City, Oklahoma

As I look out over my audience here, I see several people who really ought to be up here instead of me. I'm not an expert in the subject of Item Response Theory. Less than two years ago I had not even heard of item response theory. I discovered its existence in January of last year while thumbing through some journals. During the next several months I spent several hundred hours trying to understand it. It wasn't until last year's MTA when I was able to pick the brains of several people, that it all finally fell into place.

Soon thereafter it occurred to me that Item Response Theory really need not have been all that complicated, if someone had just sat down with me, and explained it in simple language and with a few simple examples.

The thought that all that work could have been unnecessary disturbed me to the point that I decided that no one ought to have to go through what I went through to learn about what I consider to be the most important development in the history of testing.

With that idea in mind I wrote this book. I simply put into it everything that I wish someone had told me a year and a half ago. What I intend to do today is merely to introduce some of the basic concepts of IRT. Then, if you're interested, you can get the rest of the theory from the book, hopefully.

Item Response Theory (abbreviated IRT) deals with multiple-choice questions on an ability test. But when I say "ability" I do not mean only the so-called "pure" abilities in testing, such as verbal ability, numerical ability, and spatial ability. I also mean job knowledge tests, and subject matter tests. IRT applies to all of these types of tests. It may also be applicable to personality testing, but very little work has been done on this application. It applies very well to free response (fill in) questions in addition to multiple-choice items.

Let's say we take a group of people with a wide range on some ability, say arithmetic. And let's say we give two arithmetic tests to this group, one of the tests is easy and the other is hard.

Then we will find these two distributions for the two tests.

---

*Copies of this book may be obtained from the National Technical Information Service, U.S. Dept. of Commerce, Springfield, VA 22161 by sending $8.00 for papercopy or $3.00 for microfiche. Use item # AD-A063072.

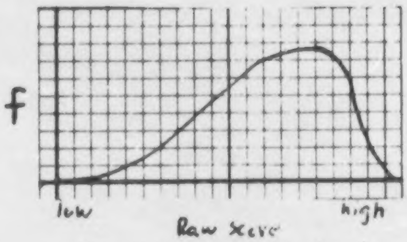


Figure 1. Easy test distribution.

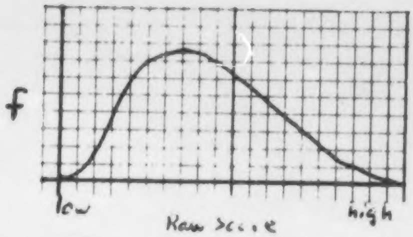


Figure 2. Hard test distribution.

The easy test will be skewed to the left because most people will score high, since it is an easy test. The hard test will be skewed to the right, because most people will score low, since it is a hard test.

In general we will find that those who score high on one test will also score high on the other test. And those who score low on one test will generally score low on the other test. And those who are at the median on the hard test will in general be at the median on the easy test. In other words, we find consistency in the performance of the examinees on the two tests of the same ability. That's not a very earthshaking observation. If we didn't find that consistency, we would not be in the testing business.

To explain this consistency we assume there is something about the examinees that causes them to score consistently relative to each other. We call that something a mental trait. No one has ever seen a mental trait and no one really expects to. Since there is no known physical referent for a mental trait, it is called a "latent" trait.

The branch of psychometrics that deals with this latent trait is called "latent trait theory".

There are several different models within latent trait theory. The models are generally distinguished by the number of parameters in the model.

There is the 1-parameter model, also known as the Rasch model. (I'll explain later what the parameters are.)

There are 2-parameter models. There are three of these. The a-b model and the b-c model, which were explored by Urry in 1970. And there is a 2-parameter polynomial model on which Samejima at the University of Tennessee is working.

The 3-parameter model is called Item Response Theory, which is the subject of this book. IRT was first presented by Fred Lord in his 1952 Ph.D. dissertation. It was called Item Characteristic Curve Theory until 1977, when Fred Lord renamed it Item Response Theory.

I'm told the Germans have an n-parameter model which means there is no limit on their number of parameters. But I know nothing about their model.

In general, there has grown a consensus that the 3-parameter model best describes reality. Most of the work in latent trait theory is now concentrating on the 3-parameter model.

Now back to the latent trait itself.

The scale of the latent trait is traditionally given the name of the Greek letter theta (θ). I will use the terms theta, ability level, amount of trait, and amount of subject-matter-knowledge, interchangeably. Theta is a continuum from minus infinity ($-\infty$) to plus infinity ($+\infty$). It has no natural zero point or unit. Therefore, the zero point and unit are often taken as the mean and standard deviation, respectively, of some reference sample of examinees. Thus, values of θ usually vary from -3 to +3, but may be observed outside that range. The θs of a sample need not be distributed normally.

When an examinee walks into a testing room, he brings with him his theta. The purpose of the test, then, is to measure the relative position of the examinees on the theta scale. The test is the measuring instrument. The test interprets the examinee's theta and produces a measurement of ability, which is often the raw (number right) score. Often measurement of an ability with a test is made analogous to measurement of height with a tape rule. But there is an important difference. Height, whether measured by an English rule or metric rule, is always on an equal interval scale. Histograms of a group of people will always look the same except for some linear stretching of a scale.

That is not the case with testing. The histograms of raw scores of the same people on two tests will seldom look the same, even with linear stretching of a scale. You can see that this is so in Figures 1 and 2. No amount of linear stretching of either scale will make the two distributions look the same. Figure 1 will always be skewed to the left, and Figure 2 will always be skewed to the right. That is because each test has its own peculiar scale (also called metric). The peculiarity of a test's metric distorts the distribution of examinees. Until IRT there has been no way to identify the peculiar scale of a test.

The traditional theory of testing is Classical Test Theory. Most testing practitioners use classical test theory, whether they know it or not. The basic tools of most testing practitioners are:

a. p-value = proportion of examinees selecting an item alternative (also called "item difficulty"),
b. d-value = point-biserial correlation between the item alternative and the test (some use the biserial correlation)(also called "item discrimination"),
c. mean of examinees' scores (number right),
d. standard deviation of examinees' scores,
e. skewness and kurtosis of examinees' scores,
f. reliability of the test, usually KR20, the Kuder-Richardson Formula 20 (a special case of Cronbach's coefficient alpha).

Anyone whose test analysis is principally based on the statistics listed above is using classical test theory. The problem with those statistics is that they are relative. They are relative to the distribution of ability among the examinees, and they are relative to the characteristics to the other items in the test.

The p-value is relative to the ability level of the examinees. The same item given to a high ability group and low ability group will get two different p-values for the two groups. It can be shown that p-values are not true measures of relative item difficulty. It is not uncommon for items measuring the same ability to reverse the order of their p-values when given to groups of different average ability. For example, item A may have a higher p-value than item B for one group of examinees, but have a lower p-value than item B for a different group. This effect is not a matter of sampling error.

The d-value is relative to the homogeneity of the ability levels of the examinees in the sample, the subject matter homogeneity of the items in the test, and the dispersion of p-values of items in the test. The same item, given to a group of examinees who are similar in ability and to another group with a wide range of ability, will produce two different d-values for the two groups. Similarly, an item included in a test with other items that are homogeneous in content and p-value will get a d-value different from the d-value it will receive in a heterogeneous test.

The mean, standard deviation, skewness and kurtosis will also vary according to the characteristics of the test and examinees.

The reliability is relative to the standard deviation of the test, and to the p-values and d-values of the items in the test, all of which are dependent upon the particular abilities of the examinees and the characteristics of the test.

It can be shown that classical parameters (e.g., p-value) will generally not be linearly related across subgroups of a population. This means that the test for cultural bias using classical parameters can lead to an artifactual detection of bias.

Clearly, classical test theory statistics are meaningful only in an extremely limited situation, i.e., when the same item is given to the identical population as part of strictly parallel tests. Such a situation rarely occurs. Furthermore, the basic precepts and definitions of classical test theory are untestable, i.e., they are tautologies. They are simply taken as true without any way to empirically determine their relevance to reality. Some are assumed to be true even when this does not appear to be warranted. Thus, no one knows if the classical test model applies to any real test.

In contrast IRT makes possible item and test statistics which are dependent neither on the characteristics of the examinees nor on the other items in the test. They are invariant. With the item statistics it becomes possible to describe in precise terms the characteristics of the test before the test is administered. This capability allows one to construct a test that is highly efficient in accomplishing the purpose of the test. It also provides an extremely powerful tool for special studies, such as item cultural bias.

Moreover, the assumptions of IRT are explicit and have the potential of empirical testing. It is possible to discover if the data reasonably meet the assumptions

The basic concept of IRT is the Item Response Function (IRF)(previously called the Item Characteristic Curve). We define 2 variables:

θ = the ability scale

$P(R/\theta) = P(\theta)$ = the probability of getting the item correct, given θ

The IRF is an S-shaped curve called an ogive (pronounced "ojive") that gives the relationship between θ and $P(\theta)$. See Figure 3.

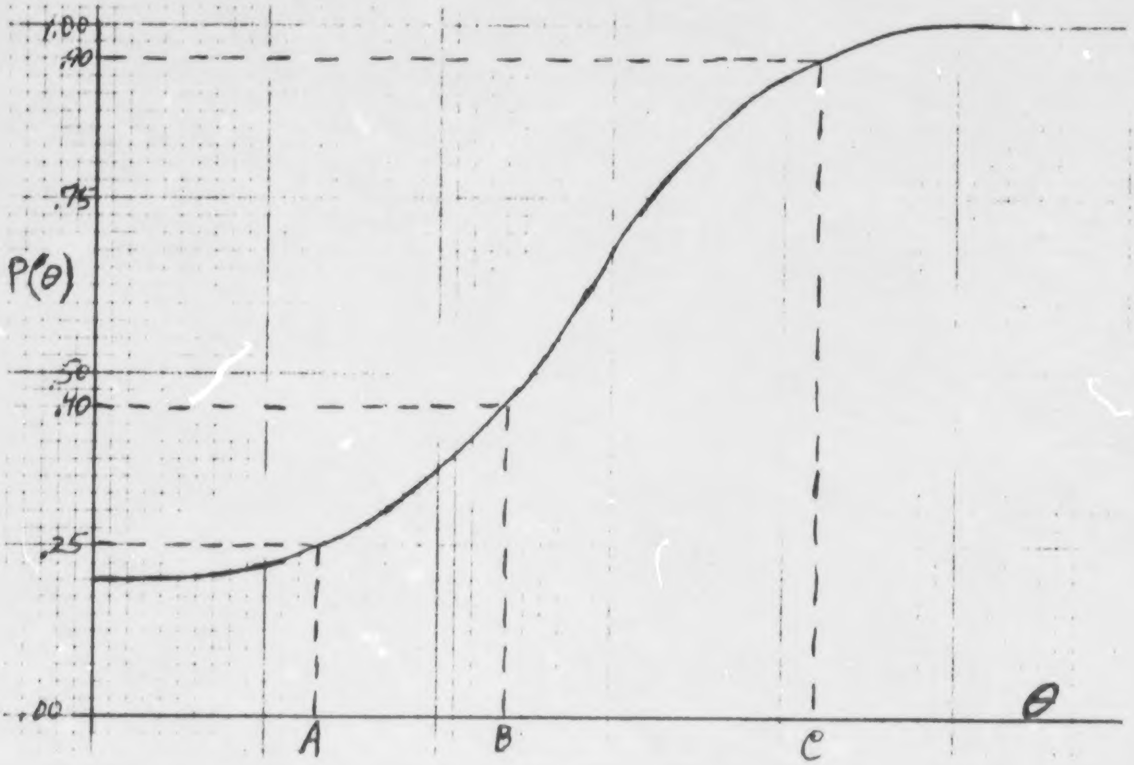


Figure 3. An Item Response Function.

Figure 3 should be read like this:

A person with the amount of ability indicated at A has a .25 probability of getting the item correct $(P(\theta) = .25)$;

A person with a θ at B has a .40 probability of getting the item correct $(P(\theta) = .40)$;

And a person with a θ at C has a $P(\theta) = .90$.

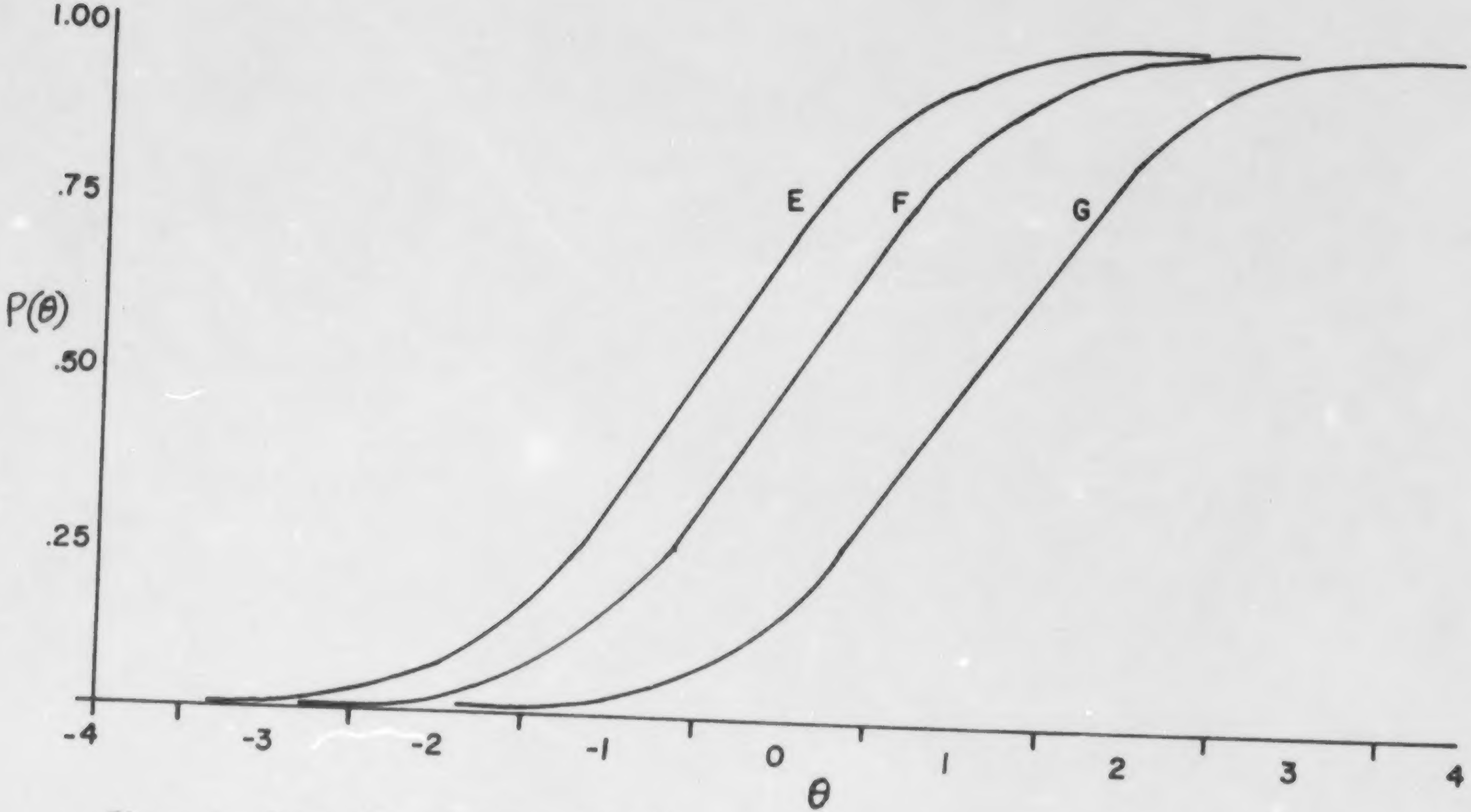


Figure 4. Three IRFs (E, F, and G) with b = -.5, 0.0, and 1.0 respectively.

Every item has its own particular IRF. Each IRF is defined by 3 parameters: the a-parameter, the b-parameter, and the c-parameter. Once these 3 parameters are known, you know everything statistically about this item that it is possible to know.

The b-parameter, or b-value as I will call it, is the horizontal location of the inflection point of the IRF. Look at the lower left part of the curve in Figure 3. That part of the curve is concave upward. The top right part of the curve is concave downward. Somewhere in the middle of the curve, it must change from being concave upward to concave downward. That point is called the inflection point. The horizontal location of the inflection point on the θ scale is the b-value of the item. The b-value is the difficulty index of the item. The larger the b-value in the positive direction, the harder is the item. The b-values of items usually vary from about -2.5 to +2.5.

Figure 4 shows the IRFs of 3 items, labeled E, F, and G, which are identical except for their b-values, -.5, 0.0, and 1.0, respectively. You can see that of the three items G (which has b = 1.0) is the hardest (i.e., has lower $P(\theta)$ for any given θ).

IRFs have 2 asymptotes. The upper asymptote is always located on the vertical axis at 1.00. In Figure 4 you can see that the upper, right part of the IRFs approach the value of 1.00 on the $P(\theta)$ axis. That is because as ability increases so does the $P(\theta)$ up to its maximum of 1.00. A probability of 1.00 is a sure thing.

The lower asymptote of the IRF is the c-value. The c-value is the probability that a person of very low ability will get the item correct.

Since we are talking about multiple-choice items, there is always a finite probability that the examinee will get the item correct by guessing.

Typically, we have assumed that the chance probability of getting the item correct is 1/A, where A = the number of alternatives in the multiple-choice question. Thus, we have assumed that a four-choice item has a c = 1/4 = .25 chance of being guessed correctly, and a 5-choice item has a c = 1/5 = .20 chance. That would be true, if examinees guessed truly randomly. But, in fact, examinees do not guess randomly when they do not know the answer. They guess according to certain patterns. Lord has suggested that item writers are very clever in writing distractors that are attractive to low ability examinees. Research has shown that when examinees do not know the answer they tend to guess the longest choice, and to avoid choices with technical or unfamiliar terms. Some examinees use a rule of thumb to always guess choice C. Whatever the reason, examinees do not guess randomly, and therefore, the c-value is seldom equal to 1/A. Typically, the c-value is .05 less than 1/A.

Most c-values range from .00 to .40. An item with a c-value of .30 or greater, is not a very good item. The lower the c-value is, the better. A c = .00 is ideal.

Figure 5 shows the IRFs of 3 items, labeled H, J, and K, which are identical except for their c-values, .30, .25, and .15, respectively. You can see that, although they all have the same b-value, they are of differing difficulty for low ability examinees.

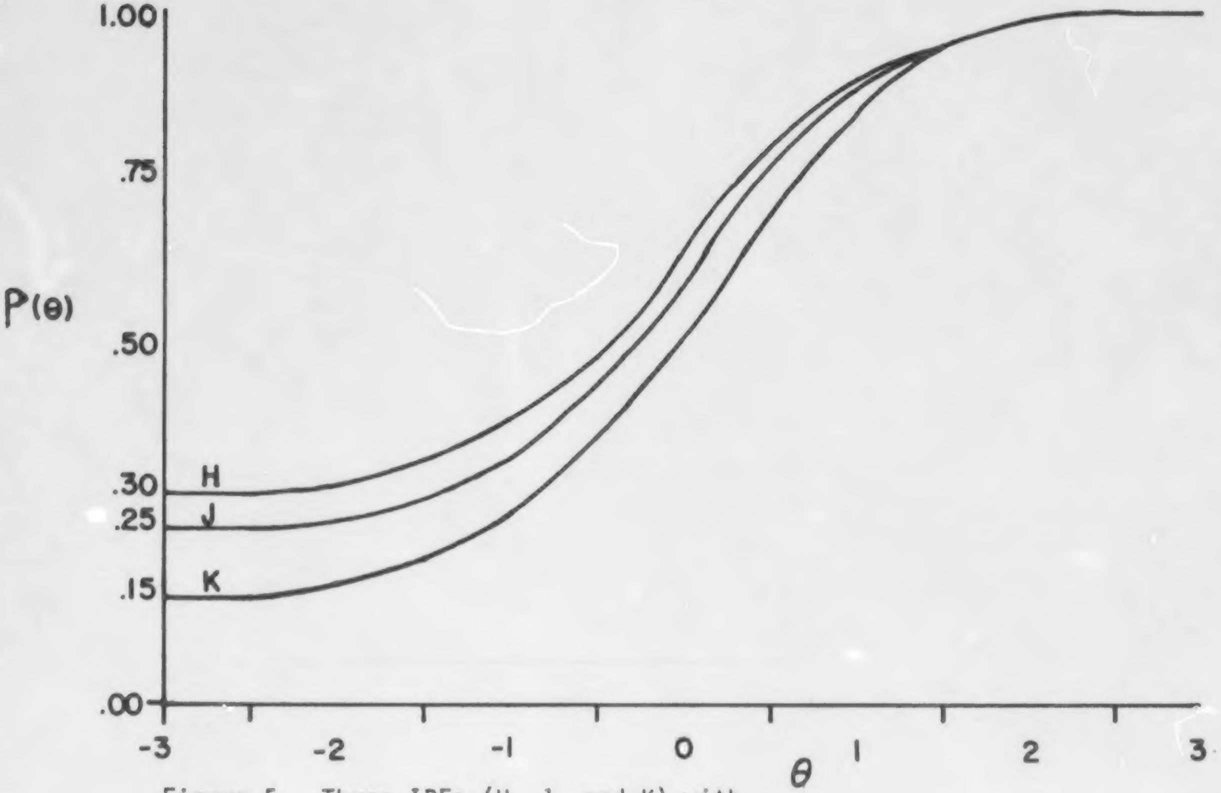


Figure 5. Three IRFs (H, J, and K) with b = 0.0 and c = .30, .25, and .15 respectively.

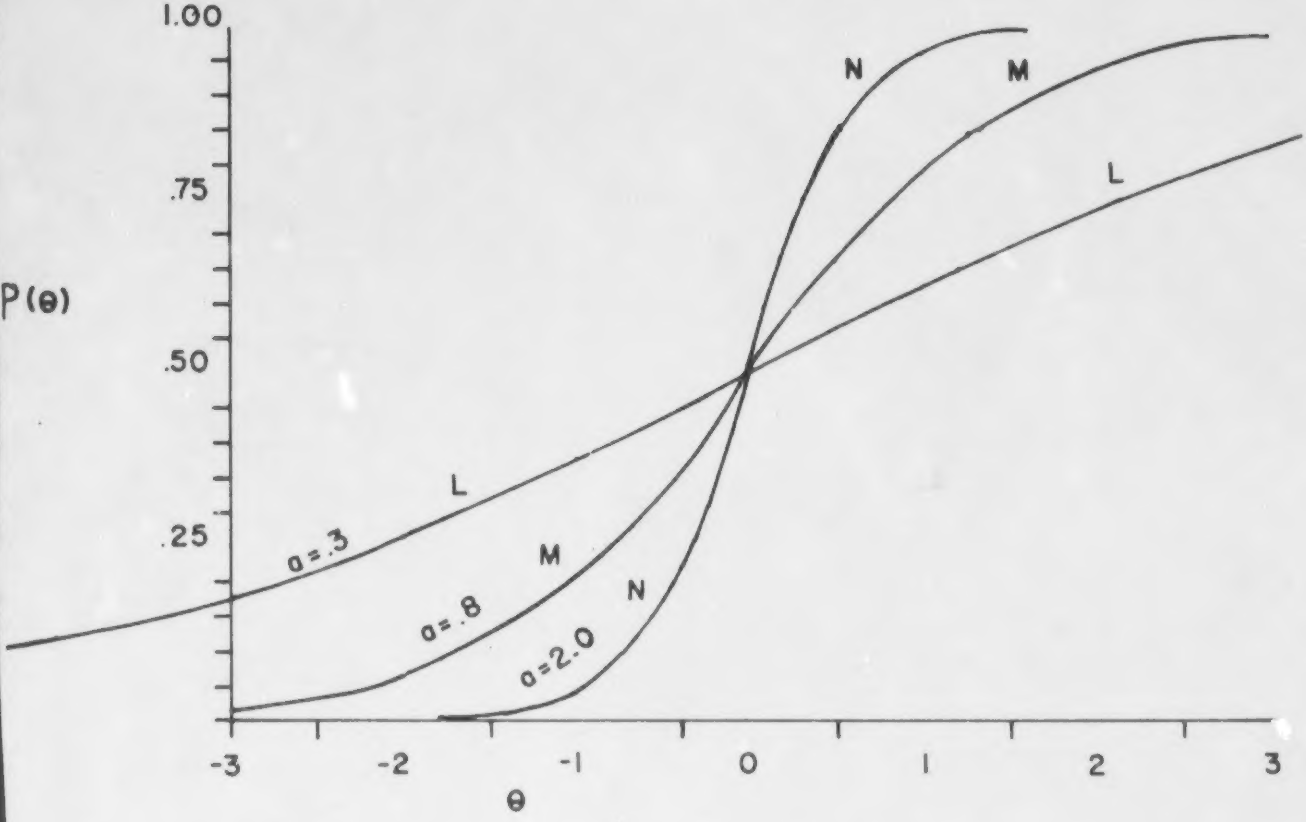


Figure 6. Three IRFs (L, M, and N) with b = 0.0, c = .00, and a = .3, .8, and 2.0 respectively.

The third (and last) parameter of IRT is the a-parameter, or a-value. The a-parameter is related to the slope of the IRT at the inflection point or in other words at the b-value. For the normal ogive model (with c = .00).

$$a = \sqrt{2\pi}\,m \approx 2.5m,$$

where m is the slope of the ogive at the b-value. Usually a-values vary from .5 to 2.5 with most between 1.00 and 2.00. The highest I have seen is 3.76.

Figure 6 shows 3 IRFs (L, M, and N), which are identical except for their a-values = .3, .8, and 2.0, respectively, with b = 0.0 and c = .00. As you can see, the larger the a-value, the steeper the IRF.

The a-value is the discrimination index of the item. The higher the a-value is, the more discriminating the item. The discriminating power of an item varies along the θ-scale. Where the slope of the IRF is high the item discriminates well. Where the slope is low the item discriminates poorly. In Figure 6 item N has high slope from θ = -1.0 to θ = +1.0, but low slope elsewhere on the θ-scale. Therefore, item N discriminates well within that range, but poorly elsewhere. A test composed of items like item N would be an excellent item for discriminating among examinees in the range θ = -1.0 to θ = +1.0. Item L has low slope across a wide range of θ. Item L discriminates a little almost everywhere on the θ-scale, but not especially well anywhere. Item L is not a very good item.

Comparing items L and N points up what is called the bandwidth paradox. You can have an item with high discrimination over a narrow range, or low discrimination over a wide range, but you can't have high discrimination over a wide range. Thus, sometimes a compromise must be made between high discrimination and the range of θ over which you have good discrimination.

Figures 7a to 7d show the IRFs of four real items from the Coast Guard Knowledge section of the Warrant Officer test.

Item #17 (Figure 7) is a hard item with high discrimination. It is the item with the highest a-value I have seen. It is an extremely unusual item for two reasons: its high a-value, and c-value equal to zero. Evidently, there is something about this item that makes nearly all examinees with θ less than +1.00 miss the item. That is a strange situation for a 4-choice item, but actually occurs for this item.

The item in Figure 7b is an easy item with somewhat low discrimination. The item in Fig. 7c is slightly easier, but has good discrimination. The item in Fig. 7d is of medium difficulty, but has poor discrimination.

Now what do you do with the IRFs once you have them? One thing you can do is to add them up. To add IRFs you merely take the height of the IRF of each of the items in a test at a particular θ-value, add them together, and plot that point. If you do this at several θ-values, and connect the points, you have what is called the Test Characteristic Curve. The Test Characteristic Curve (TCC) gives the True (number right) Score for each value of θ.

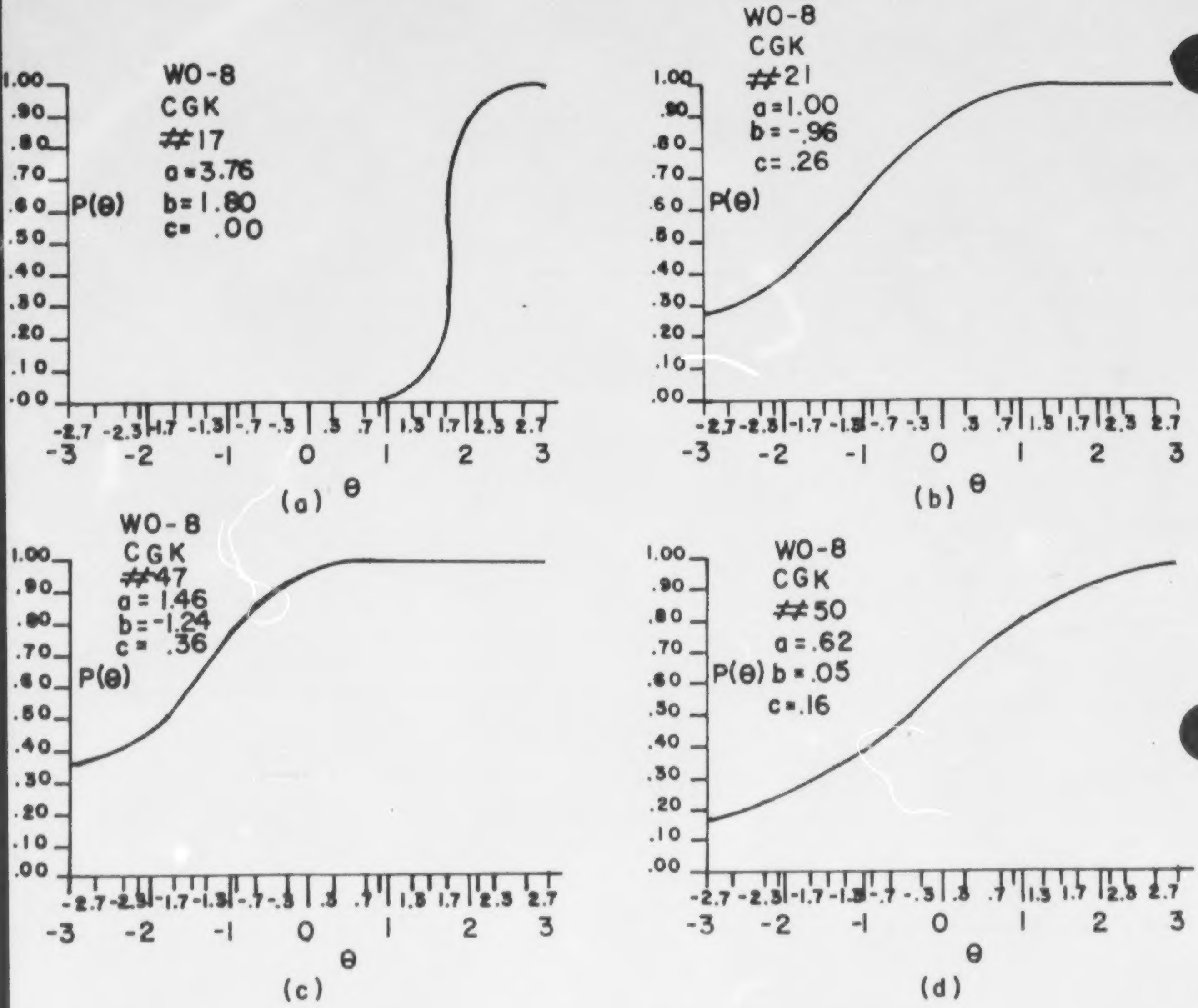


Figure 7. The IRFs of four actual items from the Coast Guard Knowledge section of the U. S. Coast Guard Warrant Officer Test, series 8.

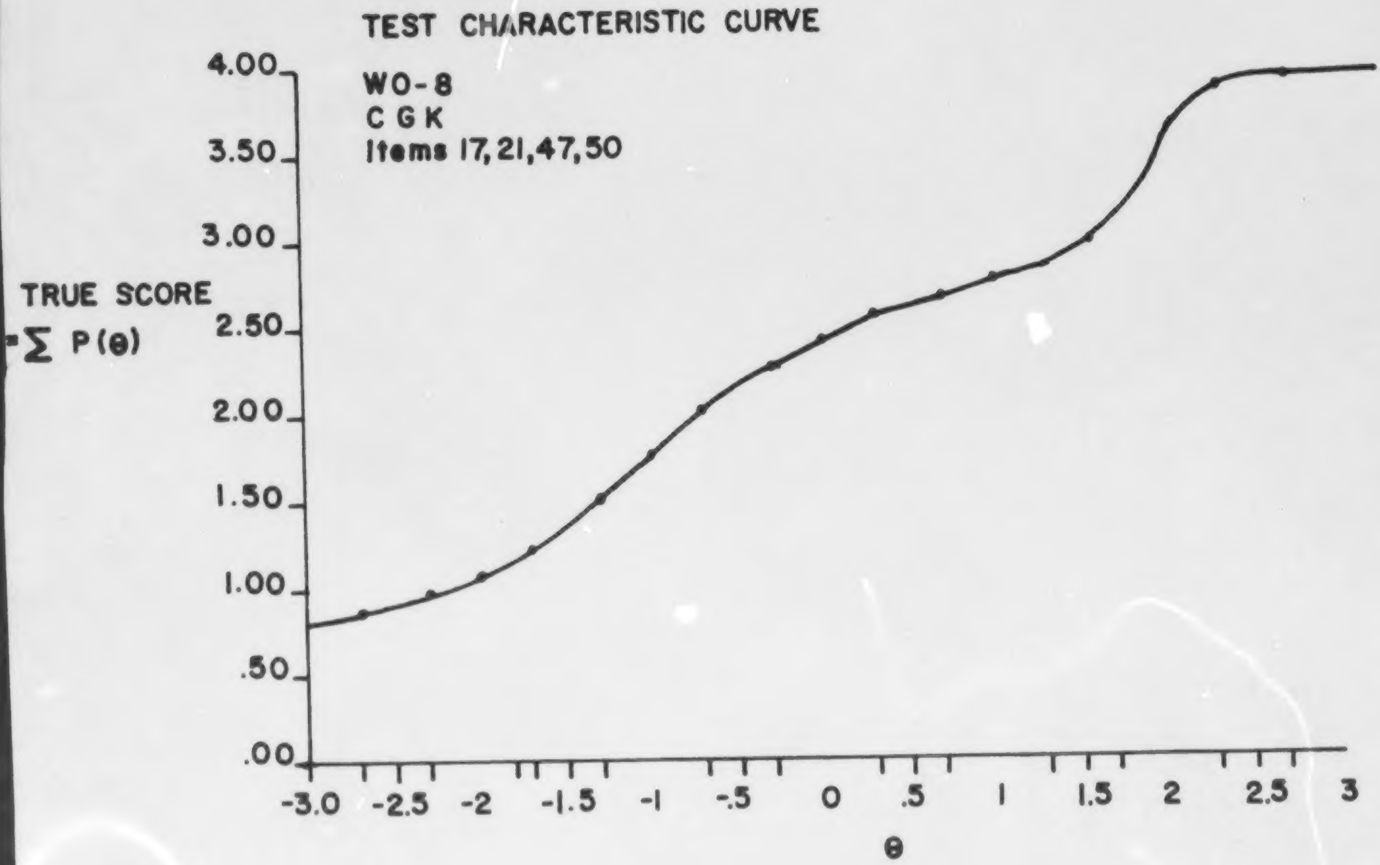


Figure 8. The Test Characteristic Curve of a test composed of four real items.

Figure 8 shows the TCC for a test composed of the four items, whose IRFs are shown in Figure 7.

Notice that the TCC is neither a straight line or an ogive. Each test will have its own TCC, which is the sum of the IRFs of the items in the test.

One of the interesting uses of the TCC is to determine the distribution of the true scores on the test. Figure 9 shows how this is done. If the examinees' θs are normally distributed, as shown on θ- (upside down), the examinees' true scores will be as shown on the left. The true score distribution is found by projecting the intervals from the θ-scale onto the TCC, and then representing the same area on the true score scale within the projected intervals. Figure 9 is an excellent demonstration of how the peculiarities of a test produce a distorted metric.

It is important to note that true scores (T) are not observed scores (X). Observed score is defined as true score plus error ($X = T + E$). However, Lord has found that the distribution of X will be similar to the distribution of T, but sometimes with the high points of the true score distribution flattened somewhat, and the low points higher. The flattening is due to error.

We can see in Figure 7a that item #17 will not help us to distinguish among examinees whose θs are less than 1.0 because they will all get the item wrong. A test made exclusively of items like #17 would do nothing to distinguish among examinees with $\theta < 1.0$ because they would all get zero on the test. It would give us no distinguishing information about them.

Item #17 also gives us no distinguishing information about examinees with $\theta = 2.7$ or greater because they will all get it correct. On a test composed of items like #17, all examinees with $\theta \geq 2.7$ would get 100%.

Between $\theta = 1.0$ and $\theta = 2.7$, it is a different story. From $\theta = 1.0$ to $\theta = 1.5$, $P(\theta)$ goes from $P(\theta = 1.0) = .00$ to $P(\theta = 1.5) = .08$. The change of $P(\theta)$ means that the item does help to distinguish among examinees within the range of θ where the change of $P(\theta)$ occurs.

We can see that the greater the slope of the IRF, the more information the item gives us about examinees in the range being considered.

The slope of the IRF would be a measure of the relative amount of information the item gives about examinees at that point. The greater the slope, the more information.

If we plot the slope of the IRF, we have a function that shows the relative amount of information an item gives at each point on the θ-scale. (Actually, the slope is not a completely appropriate measure of information, but a closely related function is.)

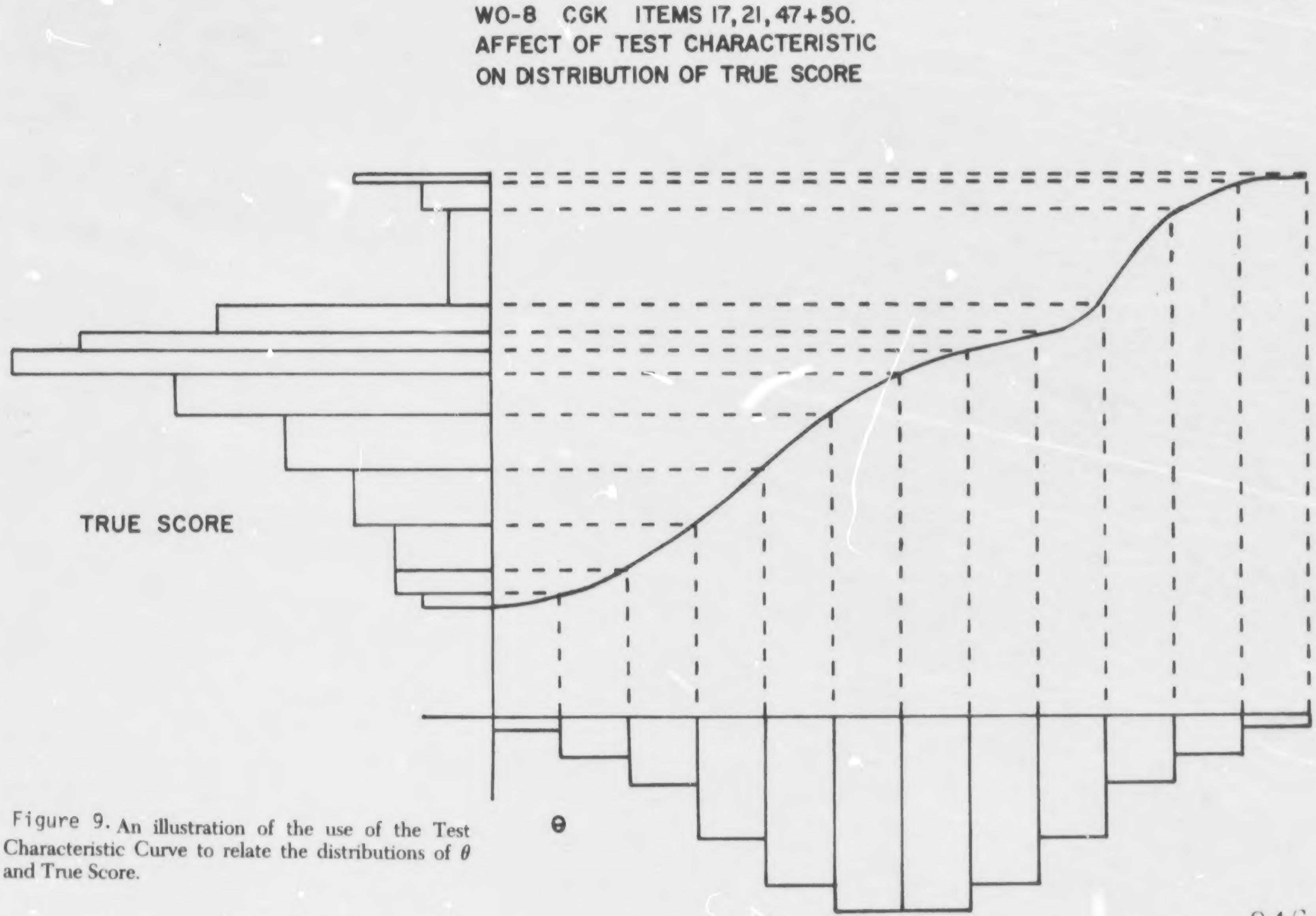


Figure 9. An illustration of the use of the Test Characteristic Curve to relate the distributions of θ and True Score.

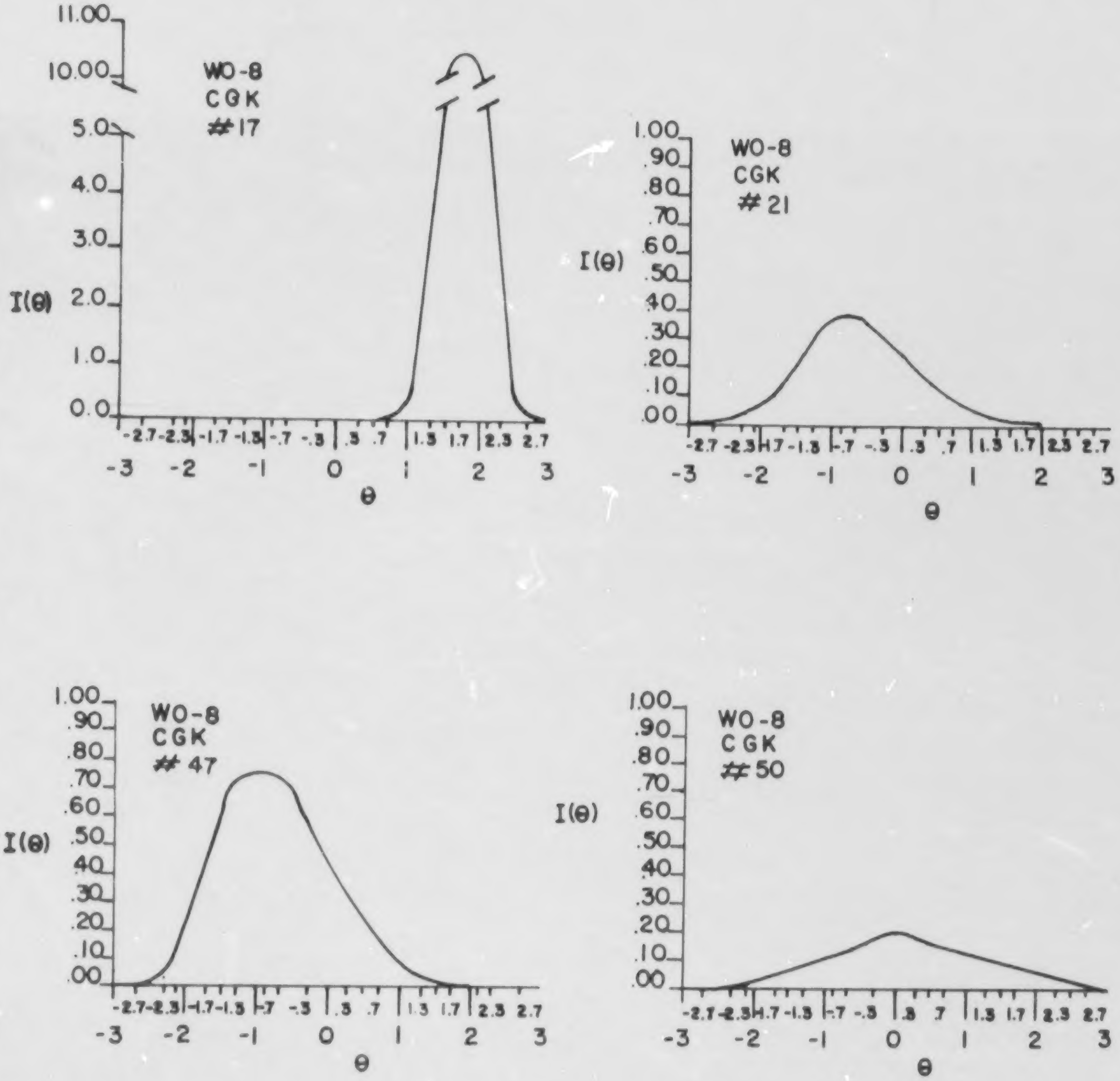


Figure 10. The Item Information Functions of four real items.

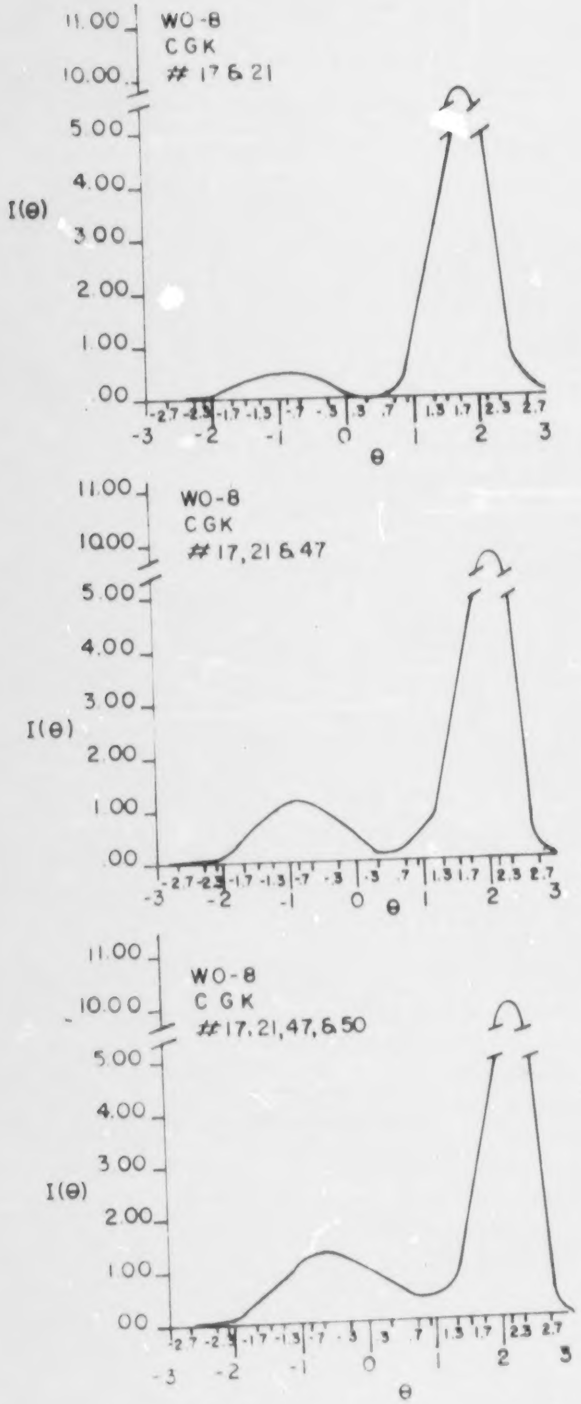


Figure 11. a, b, and c. The Test Information Curve of a test composed of items # 17 and # 21, a text composed of items # 17, # 21, and # 47, and a test composed of items # 17, # 21, # 47, and # 50 from the USCG Warrant Officer Test.

The curve showing the amount of information provided by the test along the θ-scale is called the Item Information Function (IFF).

The IIFs of the four items in Figure 7 are shown in Figure 10. (Note that the vertical axis of item #17 is a different scale from the others.) You can see the enormous amount of information provided by item #17 (with a = 3.76), compared to item #50 (with a = .62). Thus, the higher is the a-value, the more information the item provides. Also of interest is the fact that the higher is the c-value, the less information the item provides. The c-value destroys information.

What do we do with the IIF? We add them together. How do we add them together? Just like we added the IRFs together to get the TCC. We take the height of the IIFs at a particular θ-values, and connect the points. The result is the Test Information Curve (TIC).

Figure 11a shows the sum of the IIFs for items #17 and 21 as shown in Figure 10. Figure 11b shows the IIF of item #47 added to Figure 11a. Figure 11c shows the IIF of item #50 added to the other 3 items. A test composed of these four items would have the wierd TIC in Figure 11c.

The TIC shows the relative amounts of information provided by the test at each point on θ. Where you want information depends on what you will use the test for. If you want to select a few examinees from a large number, then you want a lot of information at high levels of θ, so that you can tell just which examinees are the best. For example, see Figure 12. If you want to select all examinees except a few, then you want a lot of information at low θs so you can tell which examinees are the worst (e.g., see Figure 13).

Sometimes a test is designed for more than one purpose, such as to be used with two cut scores for entrance into two different schools. In this case a two-humped TIC will give good information at the two cut scores (e.g., see Figure 14).

A TIC of any desired shape may be constructed, provided the items with the necessary IIFs are available to construct the TIC.

Usually we already have a test and want to revise it to make it better

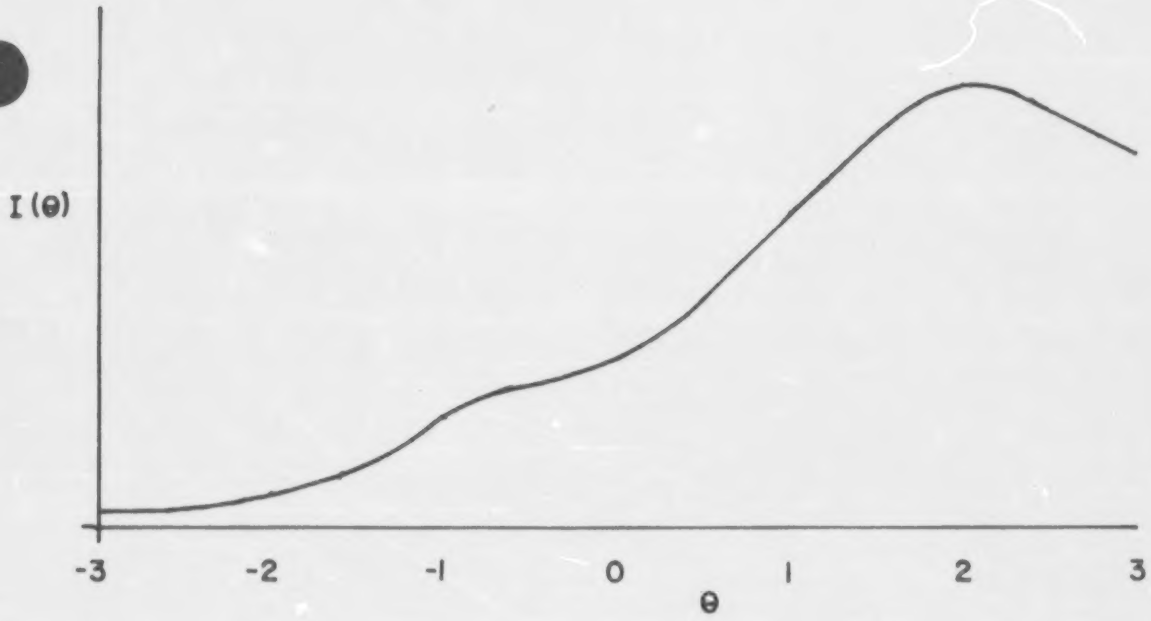


Figure 12 Test Information Curve of a hypothetical test, which would be efficient for a high cut score (Θ = 2.0).

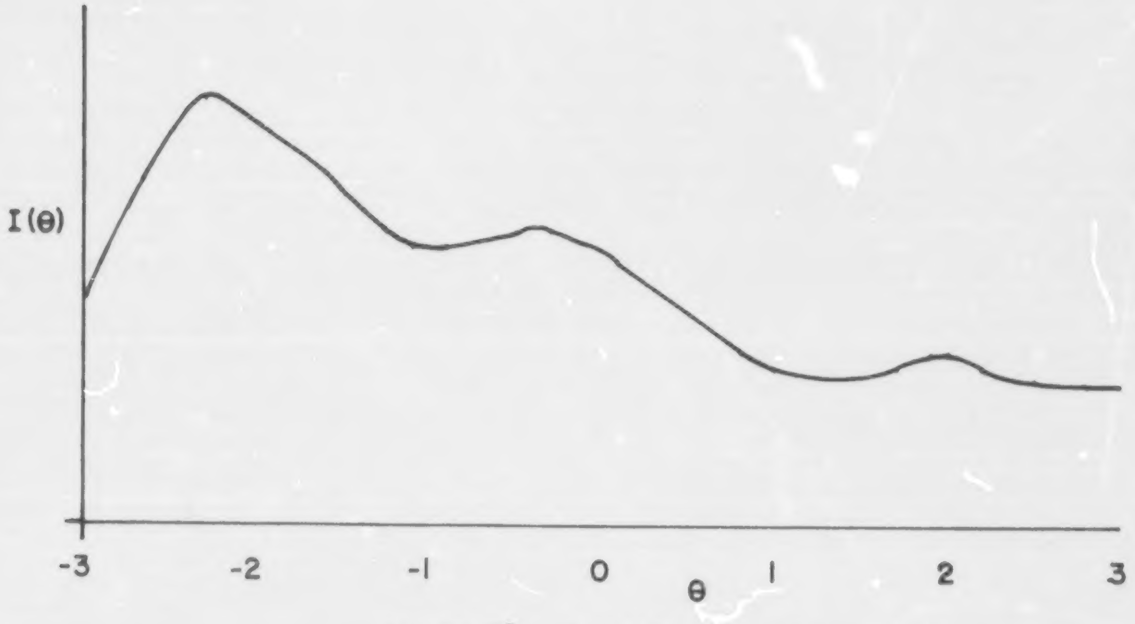


Figure 13 Test Information Curve of a hypothetical test, which would be efficient for a low cut score (Θ = -2.3).

serve our purpose. A comparison of the new and old versions should be made using the Relative Efficiency Curve (REC). The REC is nothing more than the ratio of the TICs. The ratio of the two curves is found by dividing the $I(\theta)$ of one test by the $I(\theta)$ of the other test at each point on θ. Figure 15 is the REC, comparing the TIC in Figure 14 to the TIC in Figure 13.

Where the REC is above 1.0, the test in Figure 14 (the test for which the $I(\theta)$ is the numerator of the REC ratio) is better than the test for Figure 13. Where the REC is below 1.0, the test for Figure 13 is better. And where the REC = 1.0, the two tests are the same.

By starting with an old test, making substitutions of items, and calculating the REC, you can experiment with and improve the old test by trial and error. It does not take long to develop some skill in replacing items to improve the TIC as desired.

Every test has some error in it. The Standard Error of Estimate (S.E.E.) is the expected standard deviation of errors of estimated ability. That is, if we were to give a test to a group of examinees with identical θs, and estimate their θs with the test, the standard deviation of those estimates would be the S.E.E.

If the estimate of θ is unbiased, the S.E.E. at a particular θ is easy to calculate from the TIC. The S.E.E. is equal to the square root of the reciprocal of the height of the TIC ($I(\theta)$).

$$SEE = \frac{1}{\sqrt{I(\theta)}}$$

Since $I(\theta)$ varies along the θscale, so will the S.E.E. The larger $I(\theta)$ is, the smaller the S.E.E. A small S.E.E. at a cut point is highly desirable.

The average S.E.E. ($\overline{\text{S.E.E.}}$) over examinees is related to the reliability of Classical Test Theory (r_{xx}).

$$r_{xx} = 1 - \left(\overline{SEE}\right)^2$$

This relation implies that a test with high reliability may be a poor test for your purposes because it has low information at the critical values of θ. Similarly, a test with low reliability may be an excellent test for some purposes, if it has high information where it is needed. Thus, reliability is highly misleading as to the value of a test.

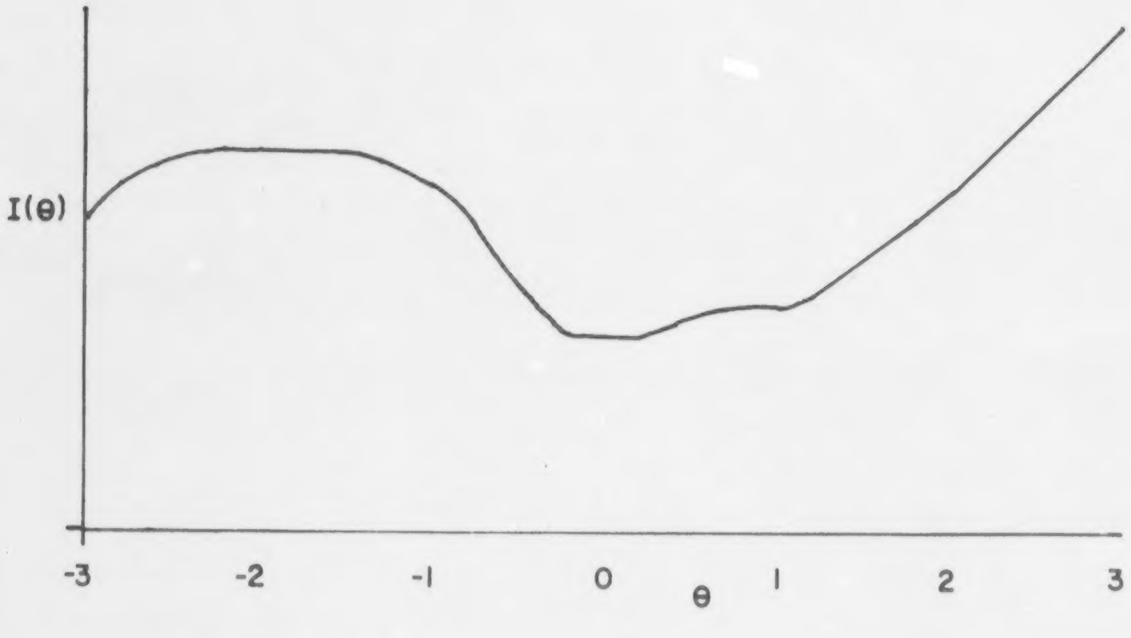


Figure 14. The Test Information Curve of a hypothetical test, which would be efficient at both high and low cut-scores.

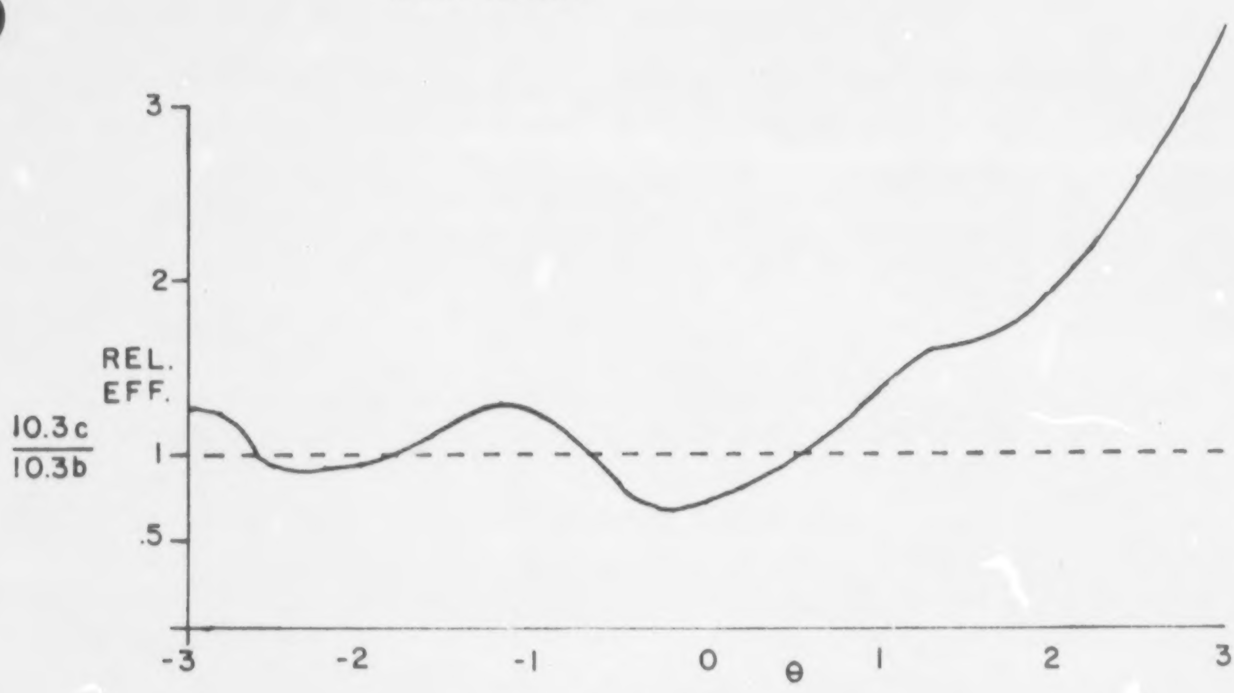


Figure 15 The Relative Efficiency Curve comparing Test Information Curve in Figure 10.3c to that in Figure 10.3b.

952

The relation also makes clear the dependence of reliability on the distribution of ability. If many examinees are on the θ scale where there is high information, then the reliability will be higher than if they are distributed on θ at points where information is low.

What are the practical applications of IRT? There are many practical applications. I will mention just a few.

First of all, IRT explains to us what a test is all about, and what an item is really doing. In my opinion, for the first time in the history of testing, testing practitioners can know what they are doing.

Second, IRT shows how to construct a test that is highly efficient for any designated purpose.

Third, IRT makes it possible to estimate an examinee's ability level with a known degree of accuracy, and without making the dubious, untestable assumptions of Classical Test Theory.

Moreover, IRT provides us with an extremely powerful tool for special studies, such as in item cultural bias.

Another exciting application of IRT is tailored testing, which is so named because it allows the "tailoring" of the test to the ability of the examinee.

Tailored tests are administered by a computer with the items presented on a CRT (Cathode Ray Tube device, which is similar to a television set). It works like this:

(1) The examinee sits in front of a CRT attached to a typewriter keyboard.
(2) The examinee registers on the computer with his identification, test name, and other pertinent information.
(3) In the computer are stored a bank of 150 to 200, or more, precalibrated items along with their item parameters. The computer selects an item of average difficulty and presents the item to the examinee on the CRT.
(4) The examinee records his answer on the typewriter keyboard.
(5) The computer uses the examinee's response and the item parameters to estimate the examinee's most likely θ, and then selects another item. The item selected is the one which will best help the computer estimate θ after the examinee answers the item. If the examinee got the item correct, he will get a different next item than if he got the item wrong.
(6) Steps (4) and (5) above are repeated until the computer meets the criterion for stopping the test.

Examinees with different response patterns will, in general, get a different set of items; yet their final estimates will be on the same metric. Not all examinees may get the same number of items, yet all θ estimates can be to the same degree of accuracy.

Tailored testing has several advantages over conventional tests.

(1) Depending upon the characteristics of the item bank, a tailored test will use only 10% to 50% of the number of items required by a conventional test and at the same time will measure more accurately than the conventional test at almost all values of θ. Tailored tests can measure to any specified degree of accuracy.

(2) A tailored test takes much less time to administer, or several abilities can be measured by a tailored test in the same time needed to measure one ability by a conventional test.

(3) Security of the items is much improved, because different examinees get different items, and because the items are much less accessible (in the computer as opposed to hard copy).

Work is progressing toward the use of tailored testing. The U.S. Civil Service Commission has adopted the use of tailored testing as a matter of policy. The U.S. Air Force Human Resources Laboratory, San Antonio, Texas, has a tailored testing machine operating on an experimental basis at the San Antonio AFEES (Armed Forces Entrance Examination Station). Several studies of live tailored testing have been published by the Psychometric Methods Program at the University of Minnesota. The Educational Testing Service is also considering tailored testing and intends to engineer its own tailored testing machine.

In closing, I hope that I have peaked your interest in IRT enough to read the entire book, where these concepts are explained in detail.

The purpose of any communication is the creation of understanding. That is my sole purpose: to create understanding of IRT in the reader.

If there is any part of this book that you do not understand, then I have not been completely successful in my effort.

Therefore, I would sincerely appreciate any comments, suggestions, corrections, ideas, or discussion about this book. Please feel free to telephone or write to me for further explanation, discussion, criticism, or just plain chew the fat about IRT.

THOMAS A. WARM, Chief, Exam Branch
Research and Examination Division
U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, OK 73169

(405)686-2417 -- commercial
732-2417 -- FTS

A NEW PROCEDURE TO MAKE MAXIMUM USE OF AVAILABLE INFORMATION WHEN CORRECTING CORRELATIONS FOR RESTRICTION IN RANGE DUE TO SELECTION

James O. Boone
Chief, Selection & Testing Research Unit
Aviation Psychology Laboratory
FAA Civil Aeromedical Institute
Oklahoma City, Oklahoma

Introduction.

To develop or update a test battery used for selecting personnel, two very important steps must be completed. First, the most valid tests must be chosen, and second, a weighting system must be devised which will combine these tests into a composite that yields a maximum validity coefficient. In order to do this all tests under consideration are intercorrelated with each other and correlated with a specified criterion of job success. These correlations are used to regress the test scores on the job success criterion and the coefficients from the regression analysis are then used to determine which tests should be included in or deleted from the battery and what the relative weights should be for each test. These weighted test scores are then combined to form the composite score which is used for selection.

In order to determine the utility of tests, both old and current tests, it is necessary to correlate them with some criterion measure of job success. Unfortunately, job success measures are available only for those individuals selected, and this selection is based on scores only on current selection tests. An important factor influencing the size of correlation coefficients between a test and the criterion is the range of scores available on the tests and on the criterion. Since information about the job success criterion is available only for applicants who have been selected for employment, only the upper range of scores is available on the criterion. Because of this restriction in range, the correlations between current selection test scores and the job success criterion will be spuriously low.

The new tests being considered to replace part or all of an existing test battery will have a larger range and variance in the selected group than the five tests actually used for selection. In fact, the range and variance will be restricted only to the extent that the new tests correlate with the old tests, and will be as restricted as the old tests only if this correlation is 1.0. Because of this differential restriction in range, the new tests will correlate higher with the job success criterion in the selected group than will the current tests.

To adjust for this spurious result, the correlations with the job success criterion must be corrected for restriction in range to assess the validity of the tests used for selection and to determine how the current tests used for selection compare with the new tests. The correction must take place prior to performance of regression analyses; otherwise, the new tests will appear superior to the current tests because of nothing more than a statistical artifact. This also means that, when corrected, the new test correlations with the criterion will generally increase less than the old test correlations.

The Uniform Guidelines on Employee Selection (1978) state that tests used for personnel selection must be demonstrated to be valid predictors of job success, and the magnitude of the validity coefficient must be both "practically and statistically significant" (3). The spuriously low correlation coefficient due to selection, then, becomes a very important legal issue in addition to its importance in assessing the value of new selection tests. Numerous litigations have occurred as a result of this problem, several of which related to the accuracy of the methods employed in correcting the validity coefficients for restriction in range (1).

There are two major statistical formulas which have been developed to correct the correlation of a test and a job success criterion. Both major formulas estimate the value of RRyz based on the information available on the restricted group: Rxy, Rxz, Ryz, Sx, Sy, and Sz. They differ in their assumptions about information available on the unrestricted group.

The first formula (5), Thorndike's formula 7 case III (hereafter referred to as T7), assumes that only SSx is available for the unrestricted group and uses the ratio SSx/Sx and the restricted correlations to estimate RRxy, RRxz, SSy, and SSz. These estimates in turn are used to estimate RRyz. The second major formula (4), Gulliksen's formula 37 (hereafter referred to as G37), assumes that only SSy is available on the unrestricted group and uses SSy-Sy and the restricted correlations and variances to estimate RRxy, RRxz, SSx, and SSz. These also are used to estimate RRyz, which is, of course, the desired unrestricted correlation of the test and the job success criterion.

The problem in using either of these formulas for the ATC selection situation is that both T7 and G37 require making estimates of either SSx or SSy and RRxy, when this unrestricted information is actually available from the applicant sample. The purpose of this study was to develop a procedure for correcting for restriction in range using available unrestricted values. In the two formulas already developed, estimates of SSz and RRxz only are required to estimate RRyz. In order to make maximum use of the unrestricted information, two formulas were derived by the author. The first formula (hereafter referred to as B1) uses SSx to derive estimates of SSz and RRxz. The second formula (hereafter referred to as B2) uses

956

SSy to derive estimates of these variables. In both formulas, the estimates, along with the actual unrestricted values of RRxy and either Sx or Sy, were used in conjunction with restricted correlations to estimate RRyz. The four formulas were compared both mathematically and by using Monte Carlo techniques to determine which can be most accurate in estimating RRyz across different selection ratios and different correlation values.

Methods.

Following Gulliksen's (4) schema for derivation of the correction formulas, three assumptions were employed, where upper case and lower case letters represent unrestricted and restricted variables respectively and x = the test used for selection, y = the new test being assessed, z = the success criterion, RR = the unrestricted correlation of the variable subscripted, SS = the unrestricted standard deviation of the variable subscripted, R = the restricted correlation of the variable subscripted, and S = the restricted standard deviation of the variable subscripted.

Employing the following assumptions;

$$Rxy\,\frac{Sy}{Sx} = RRxy\,\frac{SSy}{SSx} \qquad Rxz\,\frac{Sz}{Sx} = RRxz\,\frac{SSz}{SSx} \tag{1}$$

$$Sy^2\,(1 - Rxy^2) = SSy^2\,(1 - RRxy^2) \qquad Sz^2\,(1 - Rxz^2) = SSz^2\,(1 - RRxz^2) \tag{2}$$

and

$$\frac{Ryz - RxyRxz}{\sqrt{(1 - Rxy^2)(1 - Rxz^2)}} = \frac{RRyz - RRxyRRxz}{\sqrt{(1 - RRxy^2)(1 - RRxz^2)}} \tag{3}$$

it can be shown that

$$SSy^2 = Sy^2\left[(1 - Rxy^2) + Rxy^2\,\frac{SSx^2}{Sx^2}\right]. \tag{4}$$

and

$$SSz^2 = Sz^2\left[1 - Rxz^2 + Rxz^2\left(\frac{SSx}{Sx}\right)^2\right] \tag{5}$$

Equation (3) can be solved for RRyz, and equation (1) can be solved for RRxyRRxz to produce

$$RRyz = \frac{(Ryz - RxyRxz)\ SySz}{SSySSz} + RRxyRRxz \qquad (6)$$

and

$$RRxyRRxz = RxyRxz\ \frac{SySzSSx^2}{Sx^2SSySSz}\ . \qquad (7)$$

Substituting (7) in (6) and factoring our SySz/SSySSz,

$$RRyz = \frac{SySz}{SSySSz}\left[Ryz - RxyRxz + RxyRxz\ \frac{SSx^2}{Sx^2}\ .\right] \qquad (8)$$

Substituting the estimates for SSy (4) and SSz (5) in the root formula (8) and simplying gives:

$$RRyz = \frac{Ryz - RxyRxz + RxyRxz\ \frac{SSx^2}{Sx^2}}{\sqrt{\left(1 - Rxy^2 + Rxy^2\ \frac{SSx^2}{Sx^2}\right)\left(1 - Rxz^2 + Rxz^2\ \frac{SSx^2}{Sx^2}\right)}}\ . \qquad (9)$$

Formula (9) is equivalent to Thorndike's T7 (and also to Gulliksen's formula 19, ref. 4 p. 149).

It can also be shown from assumptions (1) through (3) that

$$SSx = Sx\ \frac{\sqrt{SSy^2 - Sy^2\ (1 - Rxy^2)}}{SyRxy}. \qquad (10)$$

and

$$SSz^2 = Sz\left[\frac{Sy^2Rxy^2 - Sy^2Rxz^2 + SSy^2Rxz^2}{Sy^2\ Rxy^2}\ .\right] \qquad (11)$$

Returning to the root equation (8), substituting the estimates for SSx (10) and SSz (11) and simplying produces the second correction formula.

$$RRyz = \frac{Rxz(SSy^2 - Sy^2) + RxyRyzSSy^2}{SSy\ \sqrt{Rxz^2(SSy^2 - Sy^2) + Sy^2Rxy^2}} \qquad (12)$$

Formula (12) is Gulliksen's formula G37.

The third and fourth correction formulas employ the assumptions of the first and second correction formulas, respectively, and make the additional assumptions that the new test under consideration, test y, was administered to the applicant group. Consequently, there is no need to estimate RRxy, SSy, or SSx, and formula (6) can be utilized as the root formula.

Substituting estimates for SSz (5) and RRxz (14) used in deriving the first correction formula (9) in the root formula (6) and simplifying gives the third correction formula,

$$RRyz = \frac{Sy(Ryz - RxyRxz)}{SSy\sqrt{(1 - Rxz^2) + \left(Rxz^2 \frac{SSx^2}{Sx^2}\right)}} + \frac{Rxz \frac{SSx}{Sx}}{\sqrt{(1 - Rxz^2) + \left(Rxz^2 \frac{SSx^2}{Sx^2}\right)}} RRxy. \qquad (13)$$

To obtain the fourth correction formula, RRxz must be derived in terms of (SSy-Sy) by first solving equation (2) for RRxz,

$$RRxz^2 = 1 - \frac{Sz^2\,(1 - Rxz^2)}{SSz^2}. \qquad (14)$$

Substituting (11) in (14), multiplying and simplifying yields,

$$RRxz = Rxz\sqrt{\frac{SSy^2 - Sy^2 + Sy^2Rxy^2}{SSy^2\,Rxz^2 - Sy^2\,Rxz^2 + Sy^2\,Rxy^2}}. \qquad (15)$$

To form the fourth correction formula, (11) and (15) are substituted in the root formula (6) and simplified giving,

$$RRyz = \frac{Sy(Ryz - RxyRxz)}{SSy\sqrt{\frac{Sy^2\,Rxy^2 - Sy^2\,Rxz^2 + SSy^2\,Rxz^2}{Sy^2\,Rxy^2}}} + RRxyRxz\sqrt{\frac{(SSy^2 - Sy^2) + Sy^2Rxy^2}{SSy^2\,Rxz^2 - Sy^2\,Rxz^2 + Sy^2\,Rxy^2}}. \qquad (16)$$

X

A demonstration of the characteristics of the four correction formulas in terms of more refined influences was performed by using Monte Carlo techniques. The Monte Carlo study examined the comparative accuracy of the four correction formulas as a function of (i) the selection ratio, (ii) RRxy, and (iii) RRyz.

In order to generate data of known means, standard deviations, and intercorrelations, a program (MNRNG) (2) was modifed by the author and used. The program uses the Marsaglia's reasonably fast method to generate normally distributed variables whose covariances are those required by a specified correlation matrix input into the program.

A summary of the process is as follows:

1. Generate 1,000 subjects with scores on 11 variables as defined by means, standard deviations, and correlations.

2. Sort sample into descending order based on scores on variable 1.

3. Restrict sample based on selection ratios of 10 percent, 20 percent, 30 percent, 40 percent, and 50 percent.

4. Calculate the four different estimates of RRyz for each restricted sample based on values of RRxy ranging from 0.2 to 0.6 and on values of RRyz ranging from 0.1 to 0.5.

5. Transform all correlations and estimated correlations by using Fischer R to Z transformation and for use in later averaging.

6. Repeat the entire process 100 times and compute the mean on the corrected correlations.

Results.

The results were prepared in tabular and graphical form. Since the sample size was 100,000, significance tests were deemed inappropriate. In order to assess the accuracy of prediction of each correlation procedure, an error term was calculated based on the absolute value of the difference between the actual unrestricted correlation RRyz and the estimated correlation Ryz. Table 1 contains this error term, RRyz - Ryz, for each correction formula, for each selection ratio, for each value of RRxy, and for each value of RRyz. Figure 1 represents this error term as a function of selection ratio for the four correction formulas and for the actual restricted correlation Ryz. Figure 2 represents the error term as a function of RRxy for the four formulas and Ryz. Figure 3 represents the error term as a function of RRyz for the four formulas and Ryz.

960

Discussion.

In any Monte Carlo study a decision must be made concerning which components are to be varied and what the range of their variation will be. The components selected for variation and their range in this study were established subjectively based on values the author considered representative of practical situations. Consequently, the discussion of the results is more a comparison of the practical utility of each formula rather than a strict mathematical comparison.

Main effects. Table 1 demonstrates the overall accuracy of each of the four formulas in terms of the average amount of error each incurred in estimating RRyz. Their rank order from least to most error is: B1, T7, B2, and G37. The first three formulas are not remarkably different; however, G37 is far less accurate than B1, T7, and B2. The clearest effect on error is produced by the selection ratio (Table 1). As the selection ratio becomes more extreme, the amount of error increases, with the increase becoming larger and larger with each step down in the selection ratio. Table 1 shows little fluctuation in error for RRxy and no systematic pattern. The effects of RRyz in Table 1 show a pattern that was found consistently throughout the analyses. When RRyz = RRxz, the error component is at a minimum. RRxz was held at a constant .30 for this study and, as can be noted in Table 1, the error increases as RRyz moves in either direction from .30.

Practical conclusions related to main effects include the following. If sufficient information is available, the B1 formula produces the most accurate estimate for RRyz. In order to have sufficient information to use B1, the new test being evaluated would need to be administered to the applicant group at the same time the old selection test is administered. Then RRxy and SSy are available for use in B1. If the new test being evaluated was not administered to the applicant group, then the most accurate correction formula would be T7 which does not require RRxy and SSy.

The selection ratio, it appears, has the largest impact on errors in estimating RRyz. If selection is extreme, 10 percent or less, the formulas for estimating RRyz are unstable and highly inaccurate. This is a difficult practical situation to resolve. A general advertisement for applicants without sufficient specific qualification statements results in a larger number of unqualified candidates and more extreme selection. However, with a highly specific advertisement self-selection becomes a secondary selection process, and the statistics computed on the applicant group are already restricted producing spuriously low validity correlations. One strategy would be to administer the selection tests to a random sample in the general population, stratifying by race and sex in order to meet Equal Employment Opportunity Commission requirements. This would yield unrestricted variances without the influence of any selection procedure.

Since RRyz is not known and RRxy is computed after the test administration, little practical guidance can be offered related to these parameters. The usual advice is clearly applicable, viz, choose a test or construct a test for selection that parallels the actual job tasks as closely as possible.

Interaction effects. As seen in Figure 1, when error in prediction is examined by selection ratio for each formula and for the actual restricted correlation of Ryz, there is a tremendous amount of error for the 10-percent selection ratio, with formula B1 doing a much better job than either T7, B2, or G37 in estimating RRyz. As the selection ratio increases beyond moderate selection (30 percent), the formulas tend to perform similarly in estimating RRyz, with the exception of G37 which consistently has more error than the other three formulas across all selection ratios.

Figure 2 demonstrates that formula B1 again is consistently the better estimator of RRyz across values of RRxy. It can also be noted from Figure 2 that as the value of RRxy increases, Ryz rapidly becomes a poorer estimator of RRyz, particularly after it passes the point at which RRyz equals RRxz (.30). Once again, G37 is a much less accurate estimator of RRyz than the other three formulas.

When RRyz is less than .30, as shown in Figure 3, B1 is the better estimator of RRyz. All formulas converge when RRyz equals RRxz (.30) and T7 is the best estimator for higher values of RRyz although the differences are small. Once again formula G37 is clearly the least accurate estimator of RRyz.

The practical implications for the interaction effects can be stated briefly. The selection ratio has such an overwhelming effect that generally the interaction effects are primarily due to the selection ratio. When the selection ratio is small to moderate (10 to 30 percent), formula B1 is clearly the most accurate estimator and should be used regardless of RRxy and RRyz. When the selection ratio goes above 30 percent, B1, T7, and B2 are practically equivalent. Formula G37 is the least desirable correction formula across conditions. Thus, overall, B1 results in the most accurate estimates of RRyz, especially when the selection ratio is 30 percent or less, regardless of the values of RRxy or RRyz.

962

References

1. Boldt, R. F.: Robustness of Range Restriction in Court. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada, 1978.

2. Boone, J. O.: Documentation for Fortran IV Programming That Produces Random Observations From a Multivariately, Normally Distributed Population Where the Means, Standard Deviations and Intercorrelations Among Variables Can Be Specified. Journal of Florida Education Association, May 1977.

3. Equal Employment Opportunity Commission: Uniform Guidelines on Employee Selection Procedures. Federal Register, 42:251, 65517, December 30, 1977.

4. Gulliksen, H.: Theory of Mental Tests. Wiley and Sons, New York, 1950.

5. Thorndike, R. L.: Personnel Selection. Wiley and Sons, New York, 1949.

Table 1. Average Error in Estimation of RRyz

Error by Formula

	B1	T7	G37	B2
Means =	0.047	0.051	0.078	0.058
Stds =	0.04	0.05	0.11	0.07

Error by Selection Ratio

	10%	20%	30%	40%	50%
Means =	0.112	0.065	0.050	0.037	0.028
Stds =	0.26	0.14	0.11	0.08	0.06

Error by RRxy

	.60	.20	.30	.40	.50
Means =	0.059	0.059	0.058	0.054	0.062
Stds =	0.16	0.14	0.14	0.14	0.18

Error by RRyz

	.10	.20	.30	.40	.50
Means =	0.073	0.055	0.048	0.060	0.056
Stds =	0.17	0.13	0.11	0.15	0.18

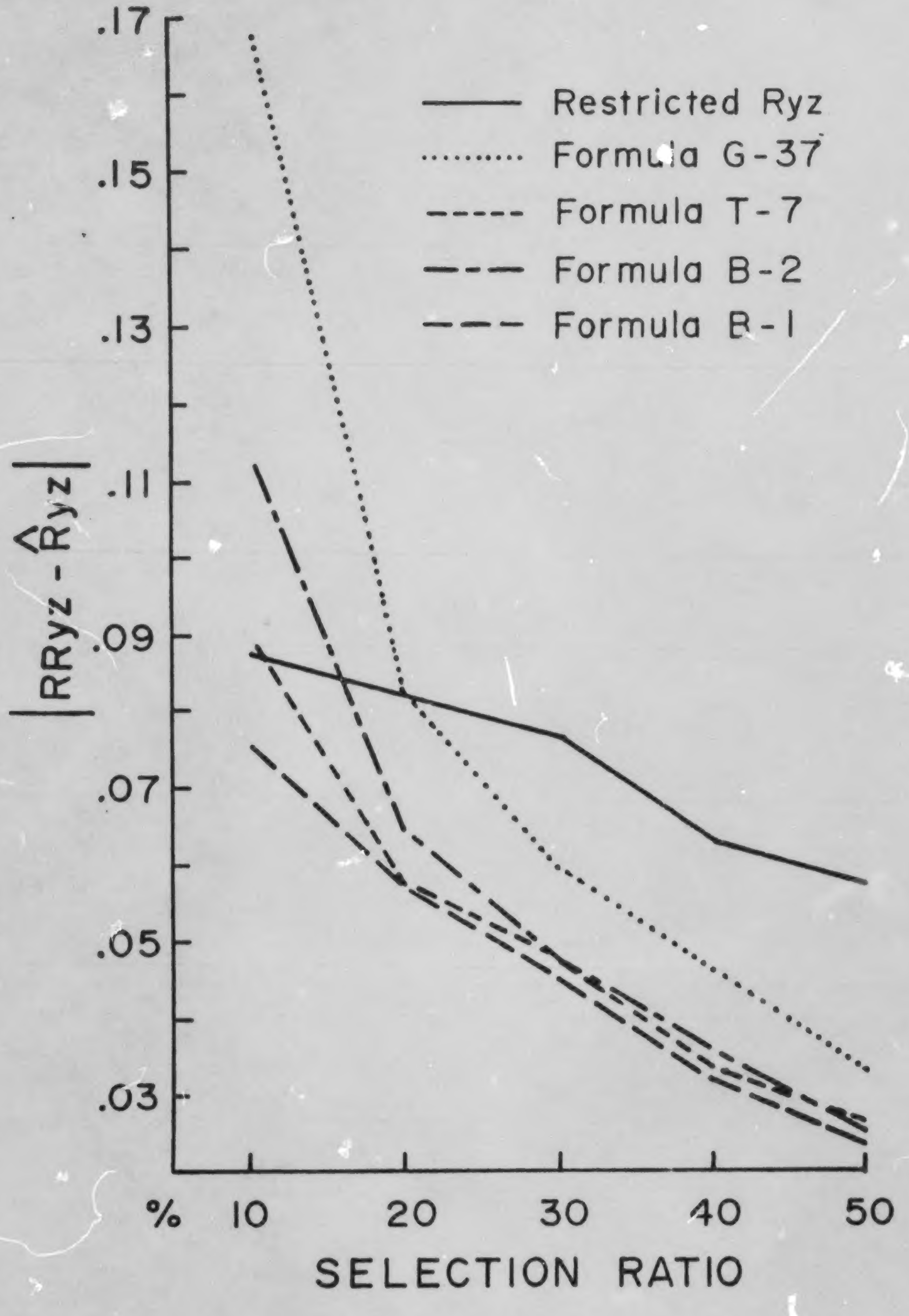
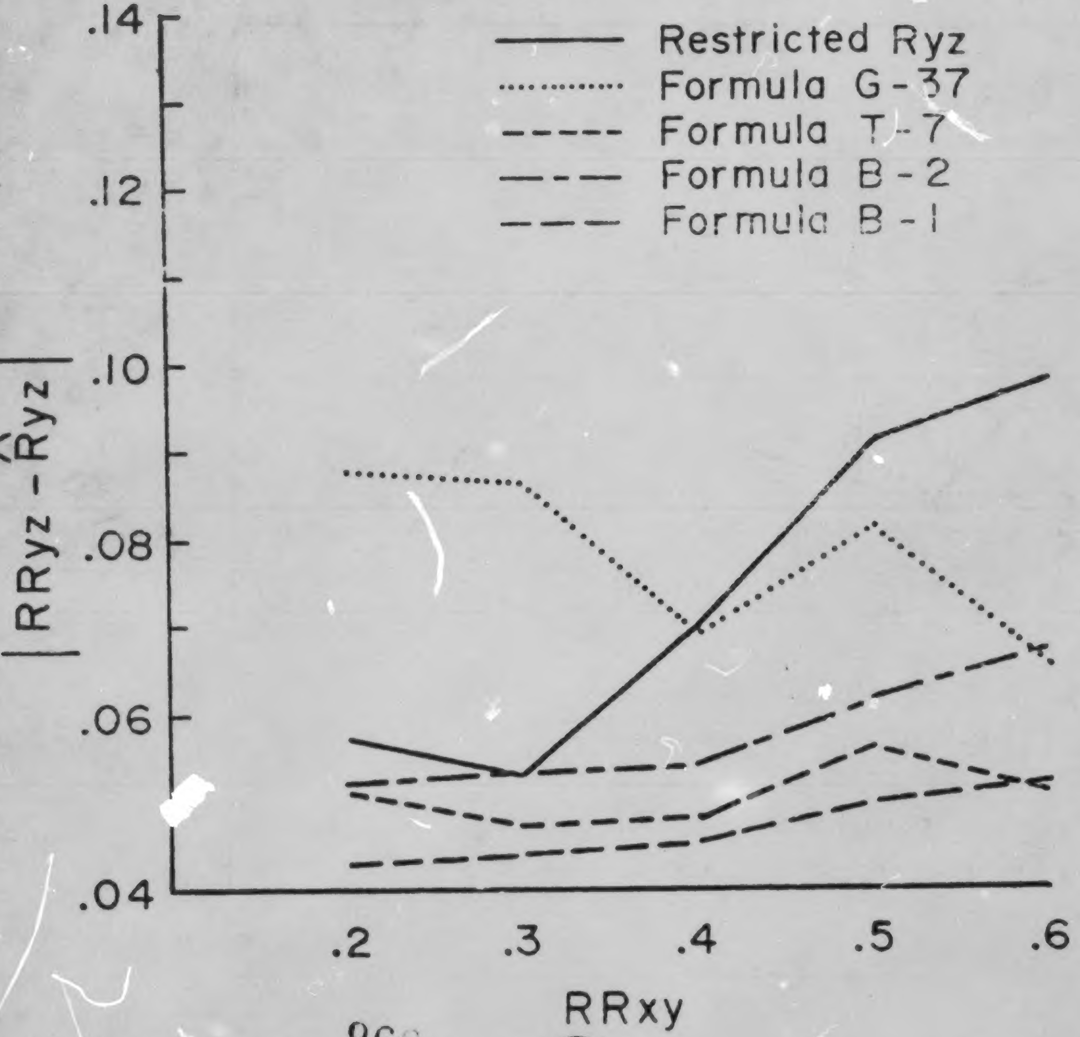


Figure 1. Error by selection ratio for the four correction formulas and the actual restricted value of Ryz.

Figure 2. Error by values of RRxy for the four correction formulas and the actual restricted value of Ryz.



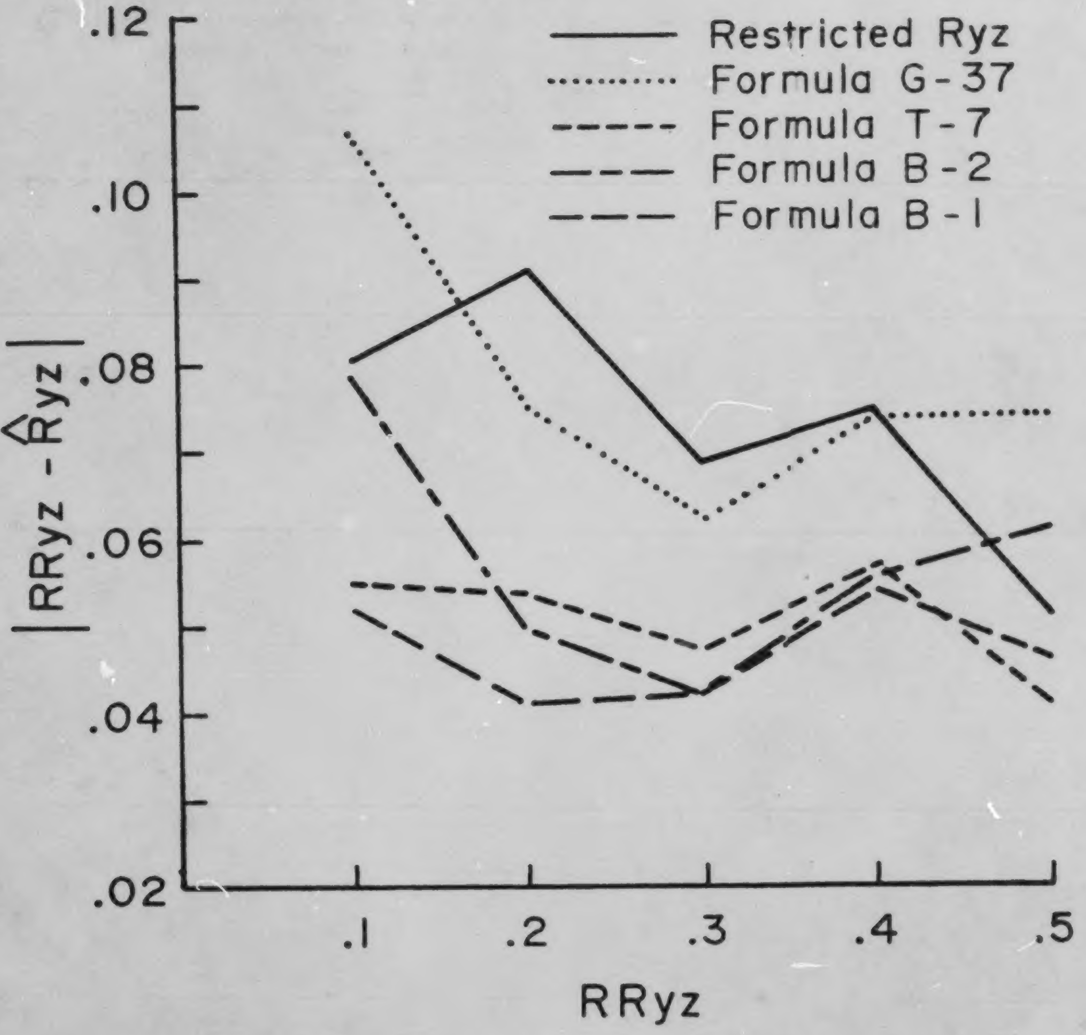


Figure 3. Error values of RRyz for the four correction formulas and the actual restricted value of Ryz.

A COMPARISON OF THREE MODELS FOR DETERMINING TEST FAIRNESS

Mary A. Lewis
Chief, Training & Evaluation Research Unit
Aviation Psychology Laboratory
FAA Civil Aeromedical Institute
Oklahoma City, Oklahoma

I. Introduction.

The Uniform Guidelines on Employee Selection Procedures (1978) (9), which were recently adopted by the U.S. Civil Service Commission, the Equal Employment Opportunity Commission, the Department of Justice, and the Department of Labor, state that a selection procedure has an adverse impact if the selection rate for any racial, ethnic, or sex group is less than four-fifths of the rate for the group with the highest selection rate. The guidelines further state that these same rules apply to any employment decision, which can include training, retention, or promotion. The current Air Traffic Control (ATC) training program conducted at the Federal Aviation Administration's (FAA) Academy is a pass/fail program which affects whether or not the trainee will be retained by the FAA in the ATC option. As such, it involves an employment decision and is subject to the standards for validation research and fairness defined by the guidelines.

Although the Uniform Guidelines acknowledge that "the concept of fairness or unfairness of selection procedures is a developing concept," they require that, when feasible, a test must be demonstrated to be fair. The guidelines further specify that "unfairness is demonstrated through a showing that members of a particular group perform better or poorer on the job than their scores on the selection procedure would indicate through comparison with how members of other groups perform." The key concept in this definition of fairness is that performance of a group is compared to the performance of the larger group on both the selection procedures and the job performance measures. If performance is not the same for both groups on both measures, unfairness may exist.

Unfortunately, deciding when "performance is not the same" is not as simple as it may seem. The literature has many articles offering approaches to the evaluation of test fairness. However, these articles seldom deal with the distribution of various fairness indices, nor do they address directly the decision processes involved in deciding whether or not a test is fair. Several authors have found that the major definitions of test fairness lead to conflicting conclusions about test fairness (1,4,7). In addition, Hunter and Schmidt (5) concede that they cannot agree on a definition of test fairness. The available literature offers many methods of evaluating test fairness but little guidance in choosing the most appropriate method.

Most of the models of test fairness define it in psychometric terms. The three major models to be discussed in the present study define fairness in the dichotomous case in which an applicant is either accepted or rejected based on a predictor score and would succeed or fail based on a criterion. Table 1 graphically depicts this situation and states the three major models of test fairness, verbally and mathematically, in terms of the four cells depicted in the table.

The first model is Thorndike's (8) Constant Ratio model (CR) which states that for a test to be fair, the ratio of the proportion successful to the proportion selected should be equal for the minority and the majority groups. Expressed in terms of the cells in Table 1, the ratio of the sum of the cells I and II to the sum of cells I and IV should be equal for both groups. Darlington's (2) Conditional Probability model (CP) states that a test is fair if the probability of selection, given that an individual is successful, is equal for both groups. In terms of the cells in Table 1, the ratio of cell I to the sum of cells I and II should be equal for both groups. Finally, Einhorn and Bass (3) propose the Equal Probability model (EP) in which a test is considered fair if the probability of success, given that an individual is selected, is equal for both the minority and the majority groups. In terms of the cells in Table 1, the ratio of cell I to the sum of cells I and IV should be equal for both groups. The three models differ in the target groups to which they are "fair." The Constant Ratio model is aimed at insuring that the proportion of applicants selected from both groups is fair. If this model is used, an equitable proportion of applicants from both groups will be hired. The Conditional Probability model is targeted at successful individuals and is intended to insure that an equitable number of successful individuals will be hired. The Equal Probability model is targeted at individuals already hired and is intended to insure that an equitable number of hired individuals will be successful. These models can lead to conflicting conclusions about the fairness of a test. However, there is very little in the literature to describe the distribution characteristics of the three models and how their distributions differ.

The purpose of the present study is to evaluate the distribution of the fairness statistics generated by the Constant Ratio, the Conditional Probability, and the Equal Probability models of test fairness. Since the sample size is, in general, much smaller for the minority sample than for the majority sample, the three fairness indices will be compared for a large sample and a smaller sample across different success ratios on both the criterion and the predictor and also across different correlations of predictor and criterion. Research studies have shown that sampling error leads to an inverse relationship between sample size and correlations (6). It is expected that sampling alone should cause the correlations for the small sample to be higher than corresponding correlations for the large sample.

Table 1. Three Definitions of Test Fairness

CRITERION		Reject	Select
	Succeed	False Negatives II	True Positives I
	Fail	III True Negatives	IV False Positives

PREDICTOR

CONSTANT RATIO MODEL (CR) - Thorndike (1971) The ratio of the proportion successful to the proportion selected should be equal for both the majority and minority groups.

$$\frac{I_a + II_a}{I_a + IV_a} = \frac{I_b + II_b}{I_b + IV_b}$$

CONDITIONAL PROBABILITY MODEL (CP) - Darlington (1971) The probability of selection, given that an individual is successful, should be equal for both the majority and minority groups.

$$\frac{I_a}{I_a + II_a} = \frac{I_b}{I_b + II_b}$$

EQUAL PROBABILITY MODEL (EP) - Einhorn and Bass (1971) The probability of success, given that an individual is selected, should be equal for both the majority and minority groups.

$$\frac{I_a}{I_a + IV_a} = \frac{I_b}{I_b + IV_b}$$

where a = majority group; b = minority group

The Constant Ratio model is not sensitive to differences in the correlation of the predictor and criterion, while the Conditional Probability and the Equal Probability models are. It is expected that the Constant Ratio model will be more robust to sampling errors related to sampling size than will either the Equal Probability or the Conditional Probability models.

II. Method.

The data used for analysis in this study were computer generated by using a Monte Carlo technique. This approach allows the generation of a number of variables with specified means, standard deviations, and intercorrelations. The technique essentially allows definition of the characteristics of a population and then selects samples from that population. A score of 70 or greater was arbitrarily set as a cut score, scores above 70 were defined as successful for the criterion variable, and scores above 70 were defined as selected for predictor. Variable means and standard deviations were assigned values such that either 60 percent, 70 percent, or 80 percent of the sample would be above the cut score, and predictor/criterion correlations of .3 or .4 were assigned. Nine variables were generated for this study by using the proportion above 70 and the correlations specified in Table 2. The success rates, selection rates, and predictor/criterion correlations were chosen based on recent experience with the FAA's Air Traffic Control selection and training program. The 18 possible combinations of selection ratio, success ratio, and predictor/criterion correlation described in Table 3 were evaluated.

Table 2. Proportion Above a Score of 70 Assigned Each Variable and Relevant Correlations Input Into Monte Carlo Program

Proportion	Var #[1]	1	2	3	4	5	6	7	8	9
.60	1	X	.3	.4	.3	X	X	X	X	X
.60	2		X	X	.4	X	X	.3	X	X
.60	3			X	X	X	X	.4	X	X
.70	4				X	.3	.4	X	X	X
.70	5					X	X	X	.3	X
.70	6						X	X	.4	X
.80	7							X	.3	.4
.80	8								X	X
.80	9									X

[1] The correlations denoted by X were not used in the analysis.

Each sample that was generated contained 1,000 subjects of which 100 were randomly assigned to the minority group and 900 were assigned to the majority group. Since both the minority and the majority groups were from the same population, the predictors should be equally fair across success ratios, selection ratios, and predictor/criterion correlations. The CP, EP, and CR indices were calculated for the 18 conditions described in Table 2. This process was repeated 100 times.

Table 3. All Possible Combinations of Selection Ratio, Success and Predictor/Criterion Correlation

	Selection Ratio	Success Ratio	Rxy	x variable	y variable
1	.60	.60	.3	1	2
2	.60	.60	.4	1	3
3	.60	.70	.3	1	4
4	.60	.70	.4	2	4
5	.60	.80	.3	2	7
6	.60	.80	.4	3	7
7	.70	.60	.3	4	1
8	.70	.60	.4	4	2
9	.70	.70	.3	4	5
10	.70	.70	.4	4	6
11	.70	.80	.3	5	8
12	.70	.80	.4	6	8
13	.80	.60	.3	7	2
14	.80	.60	.4	7	3
15	.80	.70	.3	8	5
16	.80	.70	.4	8	6
17	.80	.80	.3	7	8
18	.80	.80	.4	7	9

III. Results.

Table 4 shows the average proportion above a score of 70 and the average intercorrelation matrix obtained across the 100 large samples and the 100 small samples. Table 5 gives the distribution characteristics of three fairness indicators for both the large samples and small samples when the various combinations of selection ratios, success ratios, and predictor/criterion ratios are combined. Table 6 gives the distribution characteristics of the large and small sample fairness indicators when the selection ratio is equal to the success ratio, when the selection ratio is less than the success ratio, and when the selection ratio is greater than the success ratio. Table 7 contains the distribution characteristics of the large and small sample fairness indicator when the predictor/criterion correlation is .3 or .4.

In order to compare the fairness indices for the large and small groups, the indices were expressed first as a ratio of the large group index to the small group index (LG/SM), and then as a ratio of the small group index to the large group index (SM/LG). The distribution characteristics of these indices are described in Table 8.

Table 4. The Average Proportion Above a Score of 70 and the Average Correlation Matrix Across the 100 Large Samples and the 100 Small Samples

For 100 Large Samples[1]

Average Proportion	Var #	1	2	3	4	5	6	7	8	9
.608	1	X	0.31	0.42	0.30	X	X	X	X	X
.603	2		X	X	0.44	X	X	0.31	X	X
.643	3			X	X	X	X	0.43	X	X
.703	4				X	0.34	0.45	X	X	X
.727	5					X	X	X	0.29	X
.712	6						X	X	0.41	X
.808	7							X	0.37	0.42
.806	8								X	X
.818	9									X

For 100 Small Samples[1]

Average Proportion	Var #	1	2	3	4	5	6	7	8	9
.590	1	0.42	0.53	0.32	X	X	X	X	X	X
.583	2		X	0.30	X	X	X	0.42	X	X
.607	3			X	X	X	X	0.47	X	X
.727	4				X	0.23	0.43	X	X	X
.714	5					X	X	X	0.39	X
.700	6						X	X	0.57	X
.780	7							X	0.31	0.44
.780	8								X	X
.802	9									X

---

[1] The correlations denoted by X were not used in the analysis.

Table 5. Distribution Characteristics for the Three Fairness Indicators for the Large and Small Samples

	Mean	SD	Range	
			Lo	Hi
CRLG	1.02	.16	.74	1.35
CRSM	1.01	.18	.67	1.49
CPLG	0.77	.07	.63	0.88
CPSM	0.77	.09	.57	0.94
EPLG	0.78	.07	.63	0.88
EPSM	0.77	.09	.57	0.94

	CRLG	CRSM	CPLG	CPSM	EPLG	EPSM
CRLG	1.000	.956	- .821	- .753	.791	.737
CRSM		1.000	- .776	- .787	.758	.755
CPLG			1.000	.866	- .311	- .298
CPSM				1.000	- .324	- .202
EPLG					1.000	.902
EPSM						1.000

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample.

Table 6. Distribution Characteristics for the Three Fairness Indicators for Large and Small Samples Comparing Selection Ratio and Success Ratio

Selection Ratio Equals Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	1.017	.024	.97	1.08
CRSM	.999	.045	.89	1.11
CPLG	.773	.058	.68	.86
CPSM	.778	.076	.61	.88
EPLG	.786	.055	.69	.86
EPSM	.776	.074	.62	.89

Selection Ratio Is Less Than Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	1.194	.081	1.10	1.35
CRSM	1.220	.099	1.00	1.49
CPLG	.703	.046	.63	.77
CPSM	.698	.057	.57	.79
EPLG	.836	.035	.76	.88
EPSM	.847	.045	.73	.94

Selection Ratio is Greater Than Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	.841	.054	.74	.91
CRSM	.825	.068	.67	1.00
CPLG	.836	.035	.76	.88
CPSM	.847	.045	.73	.94
EPLG	.703	.046	.63	.77
EPSM	.698	.057	.57	.79

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample.

Table 7. Distribution Characteristics for the Three Fairness Indicators for Large and Small Samples Comparing Predictor/Criterion Correlations

Predictor/Criterion Correlation Equals .3

	Mean	SD	Range	
			Lo	Hi
CRLG	1.016	.165	.74	1.35
CRSM	1.013	.182	.67	1.49
CPLG	.761	.074	.63	.87
CPSM	.760	.088	.57	.91
EPLG	.763	.074	.63	.87
EPSM	.758	.087	.57	.91

Predictor/Criterion Correlation Equals .4

	Mean	SD	Range	
			Lo	Hi
CRLG	1.019	.145	.78	1.28
CRSM	1.017	.173	.69	1.44
CPLG	.781	.069	.68	.88
CPSM	.789	.082	.62	.94
EPLG	.787	.067	.68	.88
EPSM	.790	.081	.62	.94

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample.

Table 8. Distribution Characteristics for Ratios of the Three Fairness Indicators

	Mean	SD	Range Lo	Range Hi
CR LG/SM	1.01	.05	.88	1.15
CR SM/LG	1.00	.05	.87	1.14
CP LG/SM	1.00	.06	.86	1.20
CP SM/LG	1.00	.05	.83	1.17
EP LG/SM	1.00	.05	.86	1.20
EP SM/LG	1.00	.05	.83	1.17

	CR LG/SM	CR SM/LG	CP LG/SM	CP SM/LG	EP LG/SM	EP SM/LG
CR LG/SM	1.000	- .997	- .554	.544	.448	- .438
CR SM/LG		1.000	.574	- .563	- .426	.416
CP LG/SM			1.000	- .996	.493	- .502
CP SM/LG				1.000	- .502	.513
EP LG/SM					1.000	- .996
EP SM/LG						1.000

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample.

IV. Discussion.

As expected, Table 4 shows that the correlations for the small samples tended to be higher than those for the large samples. It is not surprising that for all three fairness indicators, the small sample groups demonstrated greater variation than did the larger sample groups. The range of the fairness indicator was virtually identical for the CP and EP models, and was a smaller range than that for the CR model. This is to be expected since the CP and EP indices could range only from 0 to 1, while the CR index could range from 0 to infinity.

When the distributions of fairness indicators are examined for the three relationships of selection ratio to success ratio described in Table 6, it can be seen that all three tend to have moderate values when selection ratios are equal; CR and EP have high values when selection ratios are greater than success ratios, while the CP value tends to be higher when the selection ratio is greater than the success ratio. Both CP and EP show the greatest amount of variance when the selection ratio is equal to the success ratio, while CR shows the greatest amount of variance when the selection ratio is less than the success ratio. When the distributions of the fairness indices for the large and small samples are examined separately for correlations of .3 and .4 (see Table 7), all three fairness indicators have lower means and higher standard deviations for the lower correlation.

The fairness indicator ratios described in Table 8 show that the distribution differences observed in Table 5 virtually disappear. The means of these ratios are around 1.00 (as they should be when the test is "fair"); the small standard deviations and the range of the ratios are almost identical for the large group/small group and for the small group/large group indices. It would appear that all three fairness indicators show similar patterns of covariance between the large sample and small sample groups.

Based on the data from the present study, there is no compelling statistical reason to choose any one of the three fairness indicators over the others. The range of the values of the indicators is affected by both the relationship of selection and success ratios, and predictor/criterion correlations. However, while the magnitude of the fairness indicator may vary, the relationship of the fairness indicators for the large and small groups remains about the same, no matter which fairness indicator is used. The three fairness indicators are equally likely to lead the investigator to conclude that a test is fair when the majority and minority groups are chosen from the same population and differences between the groups are due to sampling. Quite frequently, however, this is not the case in the real world. Members of minority and majority groups may be recruited in different ways and may differ dramatically in education, experience, socioeconomic status, and other demographic variables that will affect their performance on the selection devices. The applicants from the majority and minority groups may have different means on the selection tests, and if the means for the minority group are lower than the means for the majority group, then the proportion selected from the minority applicants could well be less than four-fifths the proportion selected from the majority applicants. If this is the case, then the Uniform Guidelines state that adverse impact has occurred, and the user must demonstrate that the selection is fair.

The Constant Ratio model could be used at this point to determine if the differential proportion selected for the minority group is compensated for by a differential success rate. If the CR definition of fairness is met, it is unlikely that the selection procedure as defined will be perceived as unfair. The CR model is insensitive to the magnitude of the correlation of the predictor and the criterion, so it would be possible to meet the CR definition of fairness while still selecting majority and minority applicants with vastly different probabilities of success. If this is the case, and if the minority group members selected have a lower probability of success than the majority group members, the minority group members will have a higher attrition rate during the training process than the majority group members. Since the Uniform Guidelines are extended to cover not just selection procedures, but also employment decisions including promotion, referral, retention, and transfer, the user may find that at some point after selection some other employment decision demonstrates adverse impact. If the Equal Probability model of test fairness is used, this problem may be avoided, but unless the regression lines for the minority and majority groups have the same slopes, its use could result in the disproportional selection of one group or the other. The Conditional Probability model could be used to insure that appropriate numbers of successful individuals are selected, but its use too could result in an inequitable selection ratio.

The test user is in a dilemma, as current definitions and practices stand. In order to meet the definition of fairness at the point of selection, the Constant Ratio model may be employed, but use of this model may result in adverse impact and unfairness at some later employment point. The acceptability of the various fairness decision models will no doubt be determined by the courts. In the ideal case, in which the minority and majority samples are selected from the same population and their regression lines are identical, all three models will agree, as they did in the present study. If the test user is in the unpleasant situation in which the models would lead to conflicting conclusions about test fairness, then some corrective action must be taken. If the Equal Probability model indicates test fairness, but the CR and CP do not, then an unfair proportion of successful minorities are being rejected, and a lower cut score may be justifiable. This will occur when the predictor criterion correlation

979

is higher for the minorities than for the majority. If the Conditional Probability model indicates test fairness, but the EP and CR do not, then the predictor/criterion correlation is lower for the minority than for the majority, and resolution of this problem may require either development of new selection procedures or recruitment of a minority applicant population that more closely resembles the majority sample.

If the use of different cut scores is not feasible, or if the data indicate that the minority applicants differ from the majority applicants in how well their performance can be predicted, the test user could examine recruitment practices to see if efforts could be made to recruit minority applicants who are more like the majority applicants in terms of characteristics related to the probability of success. The most recent version of the Uniform Guidelines emphasizes the role of recruitment and its effect on fairness. This emphasis on recruitment indicates that the effects of recruitment practices on selection and other employment decisions will be a part of the evaluation of the fairness of a selection procedure. Modification of minority recruitment practices could be an effective means of bringing existing selection procedures into compliance with the Uniform Guidelines without necessitating the development of new selection devices.

REFERENCES

1. Breland, H. M., and G. H. Ironson: Defunis Reconsidered: A Comparative Analysis of Alternative Admission Strategies, JOURNAL OF EDUCATIONAL MEASUREMENT, 13:89-99, 1976.

2. Darlington, R. B.: Another Look at "Culture Fairness," JOURNAL OF EDUCATIONAL MEASUREMENT, 8:71-82, 1971.

3. Einhorn, H. J., and A. R. Bass: Methodological Considerations Relevant to Discrimination in Employment Testing, PSYCHOLOGICAL BULLETIN, 75:261-269, 1971.

4. Hunter, J. E., and F. L. Schmidt: Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias, PSYCHOLOGICAL BULLETIN, 83:1053-1071, 1976.

5. Hunter, J. E., and F. L. Schmidt: Bias in Defining Test Bias: Reply to Darlington, PSYCHOLOGICAL BULLETIN, 85:675-676, 1978.

6. Pace, L. A., and J. L. Mendoza: Increasing Validity With Decreasing Sample Size. Presented at the Annual Meeting of the American Psychological Association, Washington, D.C., September 1976.

7. Sawyer, R. L., N. S. Cole, and J. W. Cole: Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection, JOURNAL OF EDUCATIONAL MEASUREMENT, 13:59-76, 1976.

8. Thorndike, R. L.: Concepts of Culture-Fairness, JOURNAL OF EDUCATIONAL MEASUREMENT, 8:63-70, 1971.

9. Uniform Guidelines on Employee Selection Procedures. Federal Register, Vol. 43, No. 251, p. 38290, August 25, 1978.

A METHOD TO EVALUATE PERFORMANCE RELIABILITY OF INDIVIDUAL SUBJECTS

Alan E. Jennings
Behavioral Skills Research Unit
Aviation Psychology Laboratory
FAA Civil Aeromedical Institute
Oklahoma City, Oklahoma

I. Introduction.

In laboratory research designed for eventual application to work settings, frequently the purpose is to be able to generalize performance of one population (say, college students or aviation cadets) on a complex laboratory task to a population that is highly selected for ability and motivation, e.g., airline pilots or air traffic controllers. When the tasks under consideration are complex, there is frequently a training phase of the study during which the subjects are familiarized with the tasks. If the aim of the research is to generalize to a population that is both highly skilled and motivated, it is often appropriate to select subjects during this training phase who can perform the test tasks at some minimum level of competence and who exhibit sufficient motivation to maintain consistently acceptable performance. This is especially important in this type of research because data collection is often very time consuming and costly, and practical considerations limit the sample size. An incompetent or unreliable subject can dramatically affect the accuracy of the results of such studies and, therefore, the appropriateness for applying research outcomes to the target population. An incompetent subject may be identified by specifying a minimum level of performance in the training phase of a study. However, especially in cases where repeated measure designs are employed with a small number of subjects, it would also be desirable to identify subjects who exhibit low reliability during training in order to eliminate such subjects from further training and testing. In such cases, grossly unreliable performance may be reasonably interpreted to indicate inadequate motivation or ability on the part of a subject. That is, a subject who attends to the task and performs adequately part of the time and at other times virtually ignores the task and performs at very poor levels will have corresponding variations in the task performance measure. Such variability of performance would not be likely (or acceptable) in the "real life" situations that are the ultimate concern of such research. If, for example, the researcher is generalizing to pilot performance, a pilot who was occasionally uninterested in the accuracy of his landing approach would be rapidly eliminated from the population of pilots, if not the population of the living. Thus, the elimination of subjects who clearly are able to perform adequately but who are unwilling or unable to maintain acceptable levels of performance may be an important factor in the generalizability of research findings.

In research designs where multiple measures of the same variable are made on the same subject (repeated measures), reliability of the measure is frequently estimated through the use of analysis of variance (1,4). The intent of such an estimate is to assess the stability of the test or to define homogeneous subsets of test items. The present study develops a method that may be used to estimate the reliability of an individual subject's performance across successive administrations of the same task or parallel versions of the same test and identify subjects with extremely low reliabilities. Identification of such subjects is particularly useful when the sample size is small and an unreliable subject can significantly affect the validity of the research results.

II. Method.

If, in a subjects-by-measures data matrix, all within-measure variances are equal, then the average correlation (including the diagonal) (R) among the measures is equal to the sum of squares for subjects (SS_s) divided by the quantity, total sum of squares (SS_t) minus sum of squares between measures (SS_a;

$$R = SS_s/(SS_t - SS_a).$$

If within-measure variances are unequal, then R in the above expression is a function of the sum of the covariance matrix rather than the average correlation.

This average correlation among measures (R) is an estimate of reliability of the measures, if they are parallel (6, p. 61). Parallel measures are distinct measurements that measure the same thing on the same scale (6, p. 48). Therefore, the intercorrelations of parallel measures should be equal and are the upper bound on correlations with other tests (6, p. 59).

Since the purpose of this analysis is to derive an index of subject reliability rather than measure differences, all measures must be standardized within administrations. This has the effect of equalizing the within-measure variances and results in reducing the sum of squares for measures (SS_a) to zero.

Since $SS_a = 0$, $R = \frac{SS_{subj}}{SS_{total}}$. SS_{total} is equal to the sum of SS_{subj}, and the error term SS_{ws} (sum of squares within subjects). SS_{ws} is the sum of the squared deviations of test scores around the individual subject's mean test score, which is equal to the sum of squares for the subjects-by-measures interaction.

$$SS_{total} = SS_{subj} + SS_{ws} = SS_{subj} + SS_{subj} \times a$$

R, which is used as an estimate of reliability, can then be defined as an inverse function of the within-subject variance.

$$R = 1 - SS_{ws}/SS_t$$

The within-subject variance may be calculated for any subject or group of subjects and subsequently used as an index of reliability for that subject or group of subjects.

In order to test the reliability of a given subject against the overall level of reliability, the within-subject variance for a given subject (V_i) may be compared with the within-subject variance associated with scores from the remainder of the subjects (V_{-i}). Since these two variances are independent if all subjects are independent, they may be compared by use of an F ratio. A significant V_i/V_{-i} would indicate that subject i was significantly less reliable at the specific α level than the rest of the subject sample.

The calculational procedure for these tests is as follows. Assume a data matrix X_{ij} with $i = 1$ to N subjects and $j = 1$ to M measures. These measures might reasonably be repeated measures on the same task or measures from parallel forms of the same task. The scores in the data matrix would first be standardized so that all column (measure) means and variances are equal.

Let V_i equal the within-subject variance of subject i.

$$SS_{\text{within } i} = \sum_j X^2_{ij} - (\sum_j X_{ij})^2/M \quad \text{(M = number of measures)}$$

$$df_{\text{within } i} = M - 1 \text{ so,}$$

$$V_i = SS_{\text{within } i}/df_{\text{within } i}$$

Let V_{-i} equal the within-subject variance of all subjects except i.

$$SS_{-i} = SS_{\text{within subj}} - SS_{\text{within } i}$$

$$= SS_{\text{total}} - SS_{\text{subj}} - SS_{\text{within } i}$$

$$df_{-i} = df_{\text{within subj}} - df_{\text{within } i}$$

$$= (M-1)(N-2) \quad \text{(N = number of subjects)}$$

$$V_{-i} = SS_{-i}/df_{-i}$$

Since V_i and V_{-i} are independent variances if all subjects are independent, the ratio between them is distributed as F, with (M-1) and (N-2)(M-1) degrees of freedom. A significant V_i/V_{-i} indicates that subject x is less reliable in his performance than the other subjects.

A problem in the application of this method is that it involves multiple tests, i.e., each subject is tested separately for reliability. In experimental situations where multiple comparisons are made, the Type I error rate (alpha) is much higher than the alpha level chosen for the individual tests. A straightforward solution to this problem is to use a smaller alpha value, which takes into account the number of comparisons. A simple formula (8) for the determination of alpha resulting from multiple comparisons is: $\text{alpha}_e = 1 - (1 - \text{alpha})^c$ where alpha_e is the error rate per experiment, alpha is the error rate per comparison and c is the number of independent comparisons. Although the comparisons made in the present study are not independent, this approach will identify subjects who are extreme. A table of critical values for alpha_e may be found in Jacobs (5).

In some situations, the experimenter may want to estimate the effect on R of deletion of certain subjects. This procedure is not readily amenable to significance testing but may be used to get a "feel" for the data.

R_{-i} = an estimate of the average correlation that would result if subject i were removed (assuming that for all measures, mean = 0 and s.d. = 1).

$$R_{-i} = (SS_{-i} - (\sum_j X_{ij})^2/MN/(SS_{total} - (N/(N-1)\sum_j X_{ij}^2)$$

A comparison of R and R_{-x} $(R - R_{-x})$ may be used to provide an index of the effect on overall reliability of a given subject's scores.

III. Discussion.

The method presented here provides researchers with a tool that may be used to identify subjects whose performance on repeated measures or parallel measures is unusually inconsistent. The procedure can be used for preselection of subjects for experimental studies in human factors research in which practical considerations dictate small sample sizes.

The "prediction of predictability" is a problem that has long plagued researchers (2,3,7). Using a subject reliability index as a predictability measure is a concept that has not been applied. Of course, research utilizing this method is needed to determine its potential usefulness.

References

1. Cronbach, L. J.: Coefficient Alpha and the Internal Structure of Tests, PSYCHOMETRIKA, 16:297-334, 1951.

2. Frederikson, N., and S. D. Melville: Differential Predictability in the Use of Test Scores, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 14:647-656, 1954.

3. Ghiselli, E. E.: The Prediction of Predictability, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 20:3-8, 1960.

4. Hoyt, C.: Test Reliability Estimated by Analyses of Variance, PSYCHOMETRIKA, 6:153-160, 1941.

5. Jacobs, K. W.: A Table for the Determination of Experimentwise Error Rate (Alpha) From Independent Comparisons, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 36:899-903, 1976.

6. Lord, F. M., and N. Melvin: Statistical Theories of Mental Test Scores, Reading, Massachusetts, Addison-Wesley, 1968.

7. Rock, D. A.: The Identification and Utilization of Moderator Effects in Prediction Systems, Research Bulletin 69-32, Princeton, New Jersey, Educational Testing Service, 1969.

8. Ryan, T. A.: Multiple Comparisons in Psychological Research, PSYCHOLOGICAL BULLETIN, 56:26-47, 1959.

A Comparison of Two Criterion-Referenced Scoring Procedures for An Answer-Until-Correct, Multiple-Choice Performance Test

by

John B. Meredith, Jr., Ph.D
J. Thomas Martin, Jr.
Data-Design Laboratories
Norfolk, Virginia 23502
November 1978

In many testing programs it is desirable to assess the status of the examinee with respect to a performance standard or criterion. Criterion-referenced testing (CRT) can serve as a vehicle for such an assessment. The purpose of this report is to present the results of a comparison of two CRT methods applied to a paper and pencil simulated performance test known as the Decision Measurement System (DMS).

The DMS uses a multiple-choice, answer-until-correct procedure which leads the examinee through a series of questions in an attempt to "troubleshoot" a fault within the equipment using pictorial representations of panel indications. Each examinee marks (swipes) his response on a latent image answer sheet. If the answer is correct he is directed to the next question; if his answer is incorrect he is allowed to make another swipe and continues until he has chosen the correct answer.

Two CRT methods were examined to classify examinees into pass/fail categories. The first was the present method used by the Navy. This method invoked a predetermined passing score of 62.5 for the DMS test scores, where an examinee's score is determined by exponentially combining the number of items answered correctly by him on the first, second, and third swipes.

The second method involved an extension of the Minimally Acceptable Performance Level (MAPL) technique, introduced by Nedelsky (1954) and modified by Meredith (1976), to set a passing score based on the sum of the expected number of swipes required by the Minimally Qualified Examinee (MQE) to complete each item on the DMS. The expected number of swipes for each item was determined from subject matter expert evaluations concerning the attractiveness of item alternatives to the MQE.

Each method was applied to the DMS results obtained from 30 examinees, who had been administered the Sonar Sounding Set DMS during January and February 1978.

These CRT methods were evaluated using two procedures. The first procedure was to determine the reliability of each classification method. The second procedure was to validate the CRT classifications of the examinees by determining the correlation of the pass/fail classifications with four proficiency indicators (average knowledge test scores, average skill test scores, paygrade, and number of patrols).

For each CRT method, examinees who were classified as meeting or exceeding the minimum passing score were assigned a score of one; those classified as not meeting the minimum passing score were assigned a score of zero. These dichtomous scores were used to determine the reliability and concurrent validity of both CRT methods.

The reliability of each classification method was determined by randomly splitting the DMS into two parallel sections and establishing, for both methods, a passing score on each section. Next, the proportion of consistent classifications across test sections was determined for both classification methods as an indication of their reliabilities.

The reliability of the present Navy passing score classification method was .38 while the reliability of the MAPL classification method was .64. (Note that these are conservative estimates of the classification reliabilities, since the classifications were based on half the number of original test items.) This difference in the reliability of the two classification methods was expected since the MAPL technique adapts to the difficulty of the performance test by setting a lower passing score (based on a greater number of swipes) on more difficult tests. The inflexibility of the 62.5 criterion in contrast does not account for tests of greater or lesser difficulties.

Both classification methods yielded approximately equal correlation coefficients with the four proficiency level indication. Table 1 gives the Pearson product-moment correlation coefficients for each method with the four proficiency level indicators. Neither classification method resulted in significantly higher correlation coefficients for any of the four proficiency level indicators.

The criterion-referenced MAPL technique was found to be the more efficient means for classifying examinees. Also, both classification methods were found to be equally valid when compared to four proficiency criteria. These results, however, were based on only 30 examinees, and before any sweeping generalization can be made, it is suggested that this methodology be applied to larger set of data. Further, the MAPL method for evaluating criterion-referenced performance tests should be compared to other CRT procedures, both empirically and practically.

Table 1. Pearson Product-Moment Correlation Coefficients for Classification Methods with Proficiency Level Indicators

Proficiency Level Indicators	Present Navy Classification Method	MAPL Classification Method
Average PTEP* Knowledge Score	.33	.30
Average PTEP* Skill Score	.49	.21
Paygrade	.36	.35
Number of Patrols	.27	.23

Note: For each proficiency level indicator, the difference between the correlation coefficients was nonsignificant at the .05 level.

*PTEP: Personnel and Training Evaluation Program

References

1. Hambleton, R. K. and Novick, M. R. Toward an Integration of Theory and Methods for Criterion-Referenced Tests, Journal of Educational Measurement, 1973, 10, 157-70.

2. Lord, F. M. and Novick, M. R. Theories of Mental Test Scores, Addison-Wesley Co., Reading, Mass., 1968.

3. Meredith, J. B. Determination of Minimal Acceptable Performance Levels for Criterion-Referenced, Multiple-Choice Tests, Paper presented to the 18th annual meeting of the Military Testing Association, Gulf Shores, Alabama, 1976.

4. Nedelsky, L. Absolute Grading Standards for Objective Tests, Educational and Psychological Measurement, 14: 3-19, 1954.

An analysis of the OE concept and suggested improvements

C.E. George, Henry Kinnison and H.W. Smith
Texas Tech University

We present some observations of the Army Organizational Evaluation (OE) program (USACGSC, 1978). Our concern is that the General Organizational Questionnaire (GOQ, Appendix A of TRADOC-OETC, 1974) seems to address only garrison effectiveness. If this is the case, is it possible that a tactical unit might become more effective as a garrison organization but lose some potential combat effectiveness as a result of an OE program?

A model of unit effectiveness presented earlier (George, 1977; George and Smith, 1978) suggests that this may be a real possibility (Figure 1). This model, based on experimental work with Infantry units, indicates that the GOQ factors are weak and uncertain, even potentially misleading, predictors of small unit tactical proficiency. The GOQ factors, communication flow, decision making, motivation, integration of personnel with unit and identification with unit are essentially "symptomatic" variables rather that direct determinants of tactical proficiency in small Infantry units (Figures 2 and 3).

It is recognized, or course, that the services must produce troop satisfacticn and motivation as measured by the GOQ factors. The point is that one may do this through organizational climate (higher level leadership) in ways that may fail to affect, or even degrade, tactical performance. On the other hand, it is suggested that these ends can be set on tactical problems via teamwork training.

The Army OE program

This is a voluntary program, confidentiality is promised and the anonymity of respondents is respected. The OE process consists of four steps: 1) assessment, 2) planning, 3) implementation, and 4) evaluation/follow-up. A central component of the assessment step is the 84 (plus several demographic) item GOQ. This questionnaire surveys a standard upon which to base the later steps. The items are generally easy to read and are unambiguous. They are written to fit any type of organizational setting, that is, they ask about co-workers and supervisors rather than NCO's, officers and peer-group soldiers. This generality provides some gain in adaptability but it probably also produces some feeling among combat branch, company and battalion level commanders that it is too general to fit their specific organizational concerns.

Although a commander is encouraged to work with the OE officer to add items to the GOQ, this is a laborious and uncertain procedure which may fail to produce interpretable data. It is suggested that a subset of branch specific items be developed and factor analyzed, along with the current items, on representative samples of soldiers in tactical units. At worst, this would add to the face validity of OE procedures for unit commanders. At best, it might produce a more sensitive survey instrument. It is our feeling, however, that some way must be found to measure teamwork directly in the small tactical unit and to include this evaluation as a component of the present OE program.

A specific concern we have about the GOQ is the implicit idea of the soldier as a passive recipient of information from above or a provider of information upon request. The basic requirement for developing teamwork in the lower level tactical unit, especially in the case of Infantry, is an active information seeking soldier who recognizes and meets the needs of others for data in fluid, confusing situations. It may seem unfair to criticize the GOQ for not doing something it was not designed to do. On the other hand, if these units are as different from organizations in general as we think they are, it may be vital to consider the possibility that OE users could be led to confuse garrison with operational effectiveness.

The General Organizational Questionnaire interpretive package

Computer printouts of GOQ results provide the user with highly interpretable data. Especially valuable are the breakdowns across demographic variables and by subunits of the unit being evaluated. On the negative side, the OE officer and user may be led to overinterpret the differences among subunits. Differences are said to be "moderately significant" at the .20 level. Users are warned against overinterpreting differences between medians based on small numbers of cases, but apparently not warned to take into account the total number of statistical comparisons being made.

Summary

The Army OE program in general and the GOQ specifically, have many strengths. Wider usage by lower level, combat branch commanders will probably require more believable safeguards re: confidentiality, better face (and hopefully construct) validity and perhaps further refinement of interpretive guidelines. Normative data from similar units in similar circumstances could also be most helpful.

References

George, C.E. Testing for coordination in small units. Proceedings of the military testing association, 1977, 19, 487-497 (microfiche).

George, C.E., and Smith, H.W. Satisfaction and effectiveness in the mixed gender small group. Paper presented, Psychology in the DoD symposium, USAF Academy, Co, May, 1978.

U.S. Army Command and General Staff College. Reference Book: Organizational Effectiveness. Fort Leavenworth, Kansas, 1978.

USA Organizational Effectiveness Center. General Organizational Questionnaire, Fort Ord, CA, 1977 (Including Manual and Appendices A, B,C).

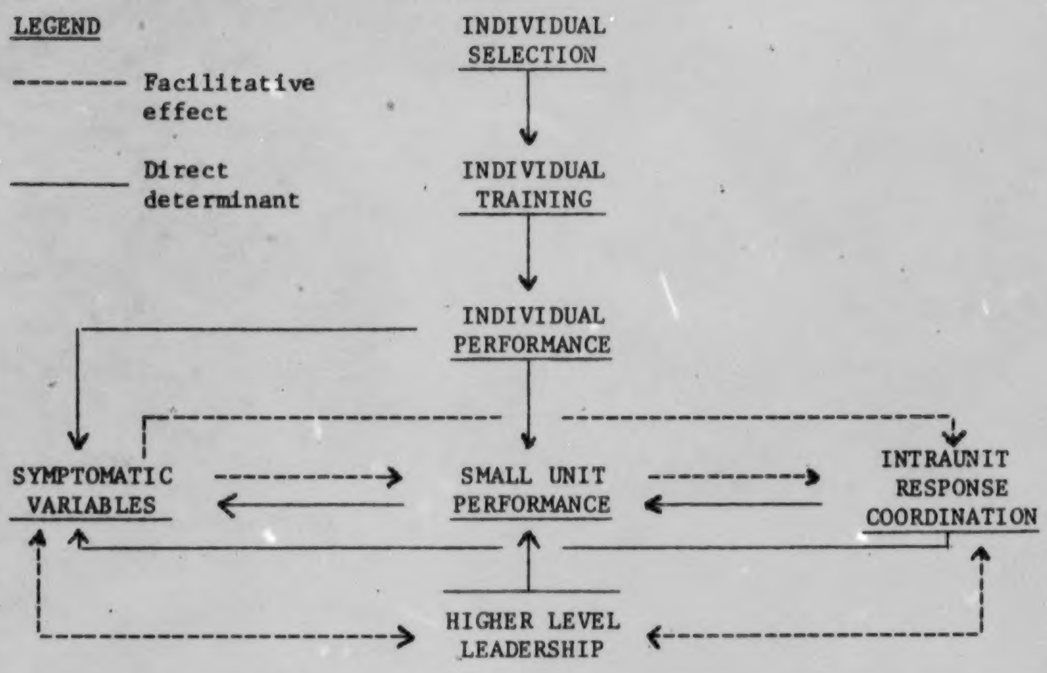
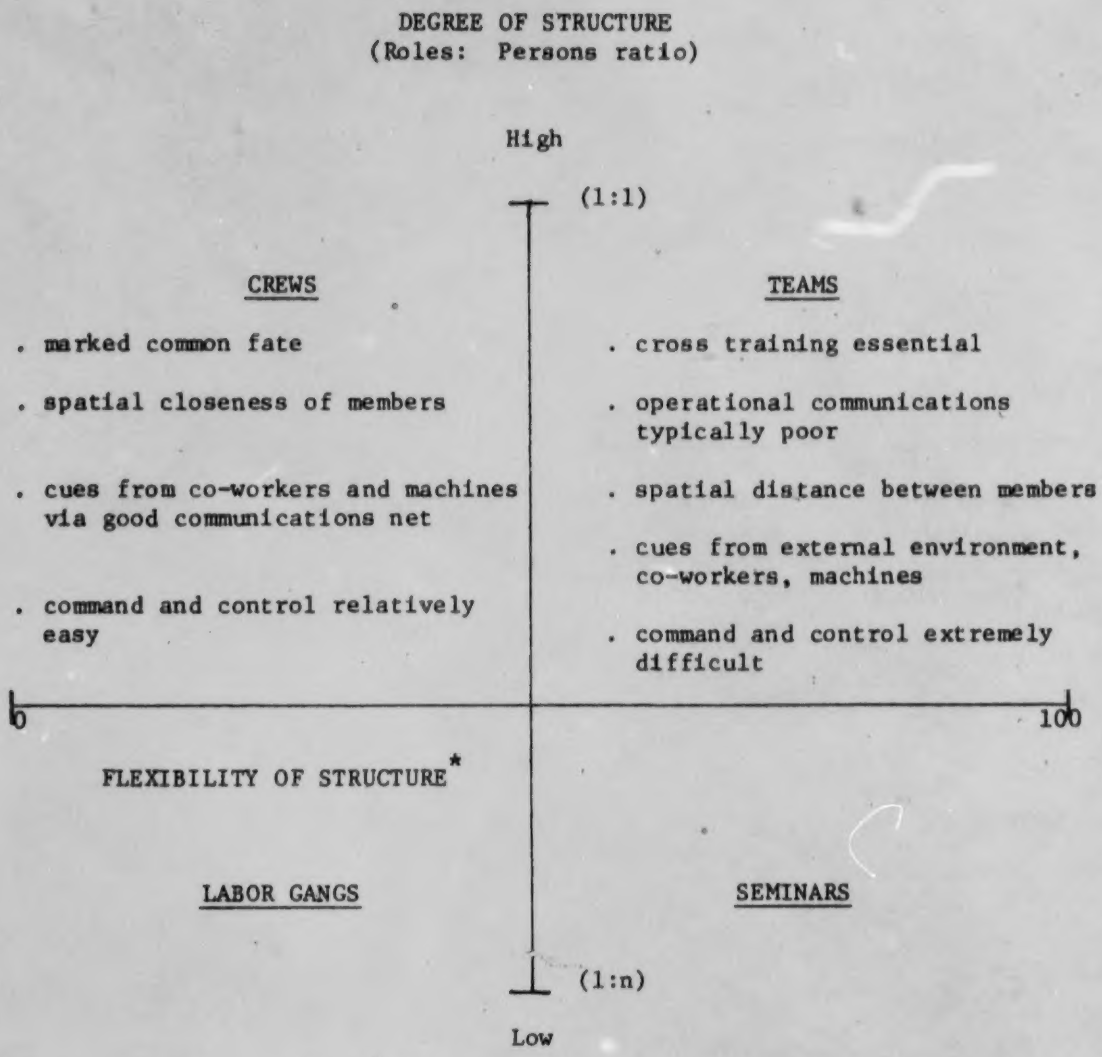

LEGEND
Facilitative effect
Direct determinant
INDIVIDUAL SELECTION
INDIVIDUAL TRAINING
INDIVIDUAL PERFORMANCE
SYMPTOMATIC VARIABLES
SMALL UNIT PERFORMANCE
INTRAUNIT RESPONSE COORDINATION
HIGHER LEVEL LEADERSHIP

Figure 1. A model of small unit functioning.


DEGREE OF STRUCTURE
(Roles: Persons ratio)
High
(1:1)
0
100
FLEXIBILITY OF STRUCTURE*
(1:n)
Low

CRECREWS

- marked common fate
- spatial closeness of members
- cues from co-workers and machines via good communications net
- command and control relatively easy

TEAMS

- cross training essential
- operational communications typically poor
- spatial distance between members
- cues from external environment, co-workers, machines
- command and control extremely difficult

LABOR GANGS

SEMINARS

*Probability of role interchange forced by uncontrollable events.

Figure 2. Model of small unit structural characteristics.

I. Symptomatic variables (individual and group characteristics within the unit-task-setting environment)
  A. Sociometric (questionable administrative utility)
    1. affection (stress resistence)
    2. respect (mutual confidence)
  B. Unit member motivation to maximize:
    1. personal achievement (intragroup competitive)
    2. socializing (emotional support)
    3. unit efficiency (coordination)

II. Behavioral coordination of response
  A. Shared attention among:
    1. one's primary job
    2. status of co-workers
    3. machine(s) in the unit system
    4. extra-unit task environment
  B. Recognition of initiative taking requirement
  C. Respond to requirement
    1. individual, immediate action
    2. communicate status to other(s)

Figure 3. Small unit level correlates of performance.

SECTION 11

TESTING: Techniques and Technologies

The Development of a Technique for Using Occupational Survey Data to Construct and Weight Computer-Derived Test Outlines for Air Force Specialty Knowledge Tests (SKTs)

by

William J. Phalen

Air Force Human Resources Laboratory
Brooks AFB, Texas

The opinions and conclusions expressed in this paper
are those of the author and are not necessarily
those of the United States Air Force.

Introduction

In June 1974, the Government Equal Employment Opportunity Coordinating Council (EEOCC) issued a slender but highly significant publication entitled Uniform Guidelines on Employee Selection Procedures. Its purpose was to spell out in some detail the intent of several acts of Congress and executive orders concerning the need to validate tests used for personnel selection. The Government specifically urged that such tests be validated against "a systematic and appropriately comprehensive analysis of the job for which the selection procedure is to be used." The Commander of the USAF Occupational Measurement Center at that time, Col Kaapke, directed that a project be initiated to establish procedures for making systematic, efficient, and timely use of job analysis data in the construction of Specialty Knowledge Tests (SKTs). This was to ensure that SKTs would be constructed in accordance with the proposals of the Uniform Guidelines. I was selected as project officer and began working on the project in September 1975.

Initial Assessment of the Problem

There had been numerous previous attempts to integrate occupational survey data into the test development process, none of which had met with much success. Early attempts had centered on the psychologist and his team of subject-matter specialists poring over the bound volumes of computer printout material that accompanied the final report of a job analysis. In most cases, the test psychologist lacked the expertise to read the printouts and locate relevant information. Even when the test psychologist was knowledgeable, the printouts themselves were not in a format that would be amenable to direct use in test outline development. Added to this, there were severe time constraints levied on the various phases of the test development process, with the inevitable result that the survey data printouts were laid aside early in the project without having made any signficiant contribution to test outline development.

Later attempts at making survey data useful for test outline development involved the preparation of a much smaller package of computer printouts which included job descriptions for SKT-relevant paygrade groupings. Oftentimes the occupational analyst responsible for the specialty would meet with the SKT team and explain how to read the printouts and how the data might be applied to the test construction process. While this procedure engendered greater use, the results still left much to be desired. The data package was still too large and complex and no systematic way was devised to make the data an integral part of the test outline development process. At best, the team would use the survey data to confirm decisions already made or to resolve debates concerning the degree to which certain tasks were performed in a specialty. But this was done after the test outline had already been developed and percentage weights had been assigned to the various content areas independently of the survey data.

If significant progress was to be made in the use of survey data in test development, it appeared that the main questions to be addressed were: Is occupational survey data relevant to the development of valid SKTs? If so, how can its relevance and usefulness be maximized?

The first question was easy to answer. Survey data had much to offer in the way of validating the content of SKTs in terms of job relevance. Because survey data are gathered on hundreds, or even thousands, of job incumbents, they provide a more representative and reliable sampling of task performance than could possibly be obtained from the judgments of three or four subject-matter specialists, no matter how broadly experienced they were. Answering the second question on how to maximize the relevance and usefulness of occupational survey data in test development is the topic that will concern the remainder of this paper. Relevance and usefulness actually subsume many other questions such as: Should all tasks in a job inventory be considered for use on an SKT, or is there only a relatively small subset of tasks in each survey that would be relevant and useful? What might be meaningful criteria for selecting relevant and useful tasks? Can a valid criterion be developed that would permit the direct evaluation of tasks on testing importance? Can survey data be used to determine not only test outline content, but also the percentage weights for outline areas? What would be the most useful format for presenting tasks for test outline and test item development? Let me now address these questions one at a time.

All Tasks vs. Subset of Tasks

Careful examination of SKT requirements and survey limitations revealed types of tasks which could be eliminated from consideration. Eight categories of task unusability for SKT purposes were identified. These, combined with a ninth "usable" category, were ordered to form a nine-point pseudo-scale that could be used by subject-matter experts to classify all the tasks in a job inventory in terms of usability. The task usability scale is shown in Figure 1.

959

Figure 1. Instructions for Coding Usability of Tasks for SKT Purposes

RECORDING TASK USABILITY FOR SKT PURPOSES

A. Rate each task in the "Time Spent" column or right-hand margin with one of the following codes (use the lowest number if more than one code applies; e.g., if codes 1, 4, and 5 apply, record "1"):

Code	Meaning
1	Task is totally inapplicable to this AFSC/shredout (if task is even slightly applicable, use code 8)
2	Task is obsolete or will soon be obsolete
3	Task statement doesn't make sense or is uninterpretable
4	Task to a large extent duplicates another task (give duty and task identifier of duplicate task; e.g., B 32)
5	Task cannot be tested by paper-and-pencil test
6	Task applies to PFE, USAF 9-Skill Level Upgrade Exam, or USAF Supervisory Exam
7	No SKT-usable reference covers this task (usually determined when attempting to write test item)
8	Task is not important enough to be tested on (e.g., very few airmen perform it, or it is extremely easy to learn, etc.)
9	Task is important enough for testing on at least one level of the AKT/SKT

B. If a task statement requires revision, record the STS area(s) and task usability code for the task statement as it is currently worded, and then pencil in the necessary revision.

C. If additional tasks need to be included in the job inventory, write in the tasks on the pages provided at the back of the inventory booklet, preceded by the appropriate duty identifier (e.g., A, B, C, etc.), and record the applicable STS areas and task usability codes for these tasks as described above.

The first four categories of unusability (1 through 4) are attributable to problems in the survey instrument. The next four categories of unusability (5 through 8) arise from special requirements of SKTs. The last category (9) is the only one which states that a task is usable. A team of subject-matter experts was asked to rate every task in the applicable job inventory on task usability for SKT purposes. Only one rating was to be given to each task: namely, the lowest-numbered rating that applied, the reason being that the lower the number, the more unusable the task. This procedure also insured that all lower-numbered categories than the one assigned to the task did not apply. The rating given was to be based on a consensus of the subject-matter experts. Separate individual ratings were not permitted, because averaging would be inappropriate for a pseudo-scale such as this one.

In general, the task usability ratings were provided by the SKT minor[1] revision team, so that the codes would be available for input to the computer in selecting tasks for a computerized test outline to be prepared for the SKT major[1] revision project the following year.

Additional requirements were also established for the selection of usable tasks. These requirements were based on task data parameters. To be selected as usable, a task had to be performed by at least 20% of one of the three groups representing the three AKT/SKT testing levels: E-2/E-3 (Apprentice Knowledge Test), E-5, and E-6/E-7. Tasks with lower percentages were excluded as unusable, as were tasks performed by a higher percentage of supervisory personnel (9-skill level) than journeyman personnel (5-skill level) and tasks which were not performed by at least 10% of job incumbents in each of the major using commands. The additional requirements were based on the fact that SKTs are Air Force-wide tests that include only speciality knowledge (no general supervision) and should cover only tasks which are performed by a significant percentage of members across the specialty (not specific to a major command). While it is true that the additional requirement aimed at the elimination of supervisory tasks overlaps code 6 of the usability scale, it has been found to be a useful backup to override coding errors.

The task filtering processes described above have routinely eliminated from SKT consideration anywhere from one-half to three-quarters of the tasks contained in the survey instrument. The remaining tasks have proved to be a quite manageable subset of tasks with strong claims to testing importance. Once the subset of usable tasks was identified, the selected tasks were assigned to one or more of the three AKT/SKT testing levels. A task had to be performed by at least 20% of the incumbents at any one level to be included as an appropriate task for testing at that level. Typically, one-fifth to one-third of the tasks would be assigned to only one level. The remaining two-thirds to four-fifths would be assigned to more than one level. So far, the number of usable tasks assigned per testing level has been between 17 and 142 tasks, depending to a large extent on the total number of tasks in the job inventory and the homogeneity of the specialty.

The Criterion Problem

As anticipated, the development of an adequate criterion to assess the testing importance of tasks proved to be the most difficult problem of all. The problem was compounded by the fact that the Specialty Knowledge Test Development Branch did not possess the resources for gathering and processing the large amount of data that would be required to obtain direct assessments of task testing importance from the more than 150 AFSCs for which reasonably current occupational survey data were available. On the other hand, obtaining task testing importance ratings from the 3- or 4-man SKT teams at the beginning of a test development project would be an exercise in futility. First of all, data supplied by such a small sample would lack representativeness and reliability, which were the major problems besetting the current method of test outline development. Secondly, the data could not be processed quickly enough to be available to the team when it was needed.

---

[1] An SKT major revision team develops new test outlines, including outline area weights, and performs a thorough rewrite of the tests. A minor revision team merely updates the previous year's test.

A different avenue which showed more promise was to use task factor data already being gathered by the Air Force Human Resources Laboratory in support of training priorities research. A factor identified as "field recommended task training emphasis" appeared to be a reasonably close analog to task testing importance--close enough that it could possibly be considered as a substitute for it. However, the substitution of training emphasis for testing importance had several drawbacks. First, the importance of a task for inclusion on a promotion test, such as the SKT, may be high, even if there is no perceived need for training in the task. Secondly, some tasks require training because they are job specific and have, therefore, not been trained in the school or encountered on previous jobs. Such tasks would be inappropriate in an SKT, which is required to test broad AFSC knowledges. Thirdly, training emphasis applies to both skills and knowledges, whereas the SKT deals only with the knowledge components of tasks. These drawbacks militated against the direct use of the recommended task training emphasis factor as the criterion of task testing importance. However, the six principal factors used by the Human Resources Laboratory to predict training emphasis seemed to encompass all the elements of testing importance as defined in the guiding documents for the SKT program. These factors were: percent of members performing the task, an index of percent time spent on the task by all members, task learning difficulty, probable consequences of inadequate performance of the task, task delay tolerance, and average grade level (by averaging the percent of members in each grade performing the task).[2] The index of percent time spent was so highly correlated with percent members performing (in excess of .90 for all observed specialties) that the index of percent time spent was dropped as being redundant. The average grade level factor was considered important as a criterion for placing tasks at the appropriate testing levels, but not as a predictor of testing importance within testing levels. The remaining four factors became the basis for two different methods of constructing a criterion of task testing importance based on a weighted composite of the four factors. The following paragraphs will be devoted to discussing the two methods.

---

[2] Evidence that these factors were relevant to the development of SKT outlines was presented in a study by Vaughan and Hickerson (reported at the 1976 MTA Conference) in which SKT test outline weights were reliably predicted from occupational data gathered on these factors.

Development of a Composite Criterion by Policy Capturing with Simulated Task Data

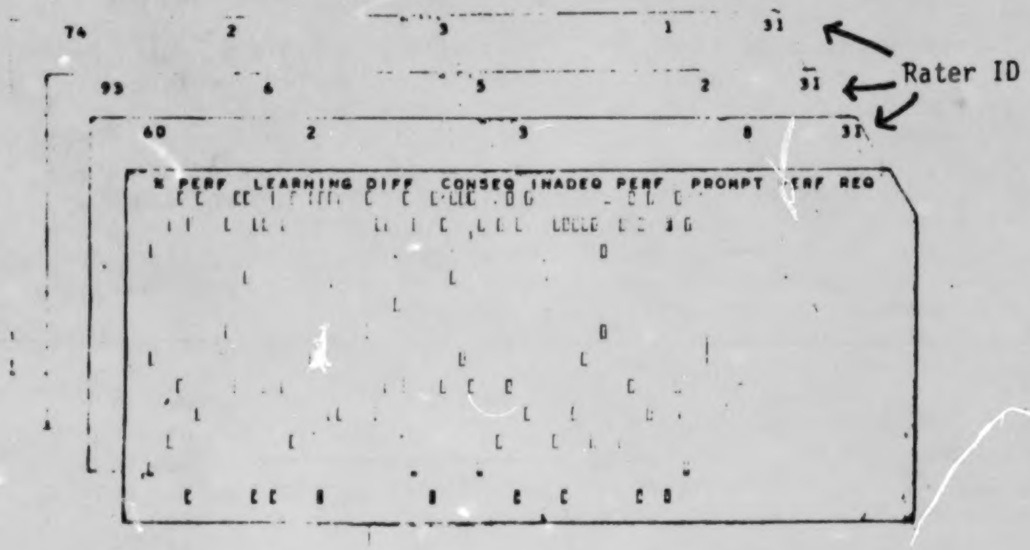
As stated previously, gathering criterion data on task testing importance from large samples of incumbents in each specialty was not feasible; on the other hand, the number of testing importance ratings that could be supplied for the tasks in each Air Force specialty by the individual SKT test development teams would not be sufficient to insure reliability.

One way of surmounting these difficulties and obtaining testing importance weights for the four predictor variables (percent performing, difficulty, consequences, and delay tolerance) based on an adequate number of raters was to develop a non-AFSC-related set of simulated tasks for which randomly generated ratings on the four predictor variables would be the only task data provided. This was done, and 56 members of 14 SKT teams were given the same set of tasks in the form of a deck of 125 randomly ordered punch cards, with each card containing four randomly generated ratings printed on the blank reverse side of each card. To avoid confusion, the task delay tolerance factor, which used a reversed scale relative to testing importance (1=least tolerance for delay, 9=most tolerance for delay) was reversed and called "requirement for prompt performance" to make it directionally comparable to the other three factor scales. A blank card containing factor titles was also furnished so that the ratings could be identified with the appropriate factor by superimposing the factor titles card on the data card. Figure 2 shows three simulated task cards and a factor titles card.

Each subject-matter specialist was asked to rankorder the cards (tasks) on testing importance using the information provided on the four factors. It was up to the subject-matter specialist to visualize what kind of task might fit the data on each card. The actual ranking of the cards was performed only after the cards had twice been sorted into five categories of testing importance (5 x 5 = 25 categories) in order to simplify the ranking process. A regression equation was computed for each subject-matter specialist using the testing importance rankings as the criterion variable against which to regress the ratings on the four predictor variables. Four cases from one SKT team were dropped because the members apparently did not perform the rankordering, as evidenced by the extremely low correlations of all four predictor variables with the criterion for those cases. Two other cases were dropped because of missing cards.

A hierarchical clustering of the regression equations of the remaining 50 cases was performed to determine whether there was more than one ranking policy employed by the subject-matter specialists. Four distinct ranking policies were identified, as shown in Table 1.

Figure 2. Factor Titles Card and Three Simulated Task Cards


74
2
3
1
31
93
6
5
2
31
60
2
3
8
31
Rater ID
% PERF LEARNING DIFF CONSEQ INADEQ PERF PROMPT PERF REQ

NOTE: Punched holes in card represent alphanumeric characters used in titles.

Table 1
Beta Weights and R^2 Values for Four Testing Importance Policies Identified by Hierarchical Clustering of Regression Equations, Using Simulated Tasks

		Beta Weights				
POLICY	N	% PERF	LRN DIFF	CONSEQ	DELAY TOL*	R^2
A	15	.3073	.1033	.4495	.4707	.5736
B	23	.0318	.1144	.8856	.1177	.8047
C	5	.0124	.7965	.1504	.0909	.5634
D	3	.9595	.0232	.1072	.0102	.9174
OVERALL	50	.1715	.1650	.5622	.1896	.4153

*Task delay tolerance scale was reversed and called "Requirement for Prompt Performance."

Differences among policies are significant beyond .001 level of confidence.

Policy A, which was used by 15 cases, gave the greatest weight to consequences of inadequate performance and task delay tolerance, and also gave substantial weight to percent members performing. Policy B, which was used by 23 cases, gave overwhelming weight to consequences of inadequate performance. Policy C, which was used by five cases, gave overwhelming weight to task difficulty. Policy D, which was used by three cases, gave overwhelming weight to percent members performing. Four cases did not fall into any policy group. There did not appear to be any identifiable AFSC pattern associated with the policy groupings such that different equations could be called upon in comparing task testing importance values for individual specialties or groups of related specialties. Intercorrelations of the predictor variables are not reported here, for the obvious reason that the random assignment of factor ratings to the simulated tasks insured virtually zero correlations between these variables. While it is true that the data which the Human Resources Laboratory has gathered on the four predictor variables show substantial intercorrelation, the goal of this policy-capturing effort was to obtain uncontaminated correlations of the factors with the criterion (testing importance). If intercorrelation had been built into the assignment of ratings to the simulated tasks, the fact that the four ratings on each punch card were locked together in the ranking process would have induced spurious correlation of each factor with the criterion.

Once uncontaminated correlations of the predictor factors with the criterion were obtained, the real intercorrelations of the predictor variables, which differed from one specialty to another, were plugged into an appropriate regression model to compute task testing importance values using the four-factor composite. Results and comparisons of this method with the second method will be discussed after the second method of criterion development has been presented.

Development of Composite Criterion Based on Factor Importance Ratings

Upon completion of the rankordering of the simulated tasks, each subject-matter specialist was asked to rate on a nine-point scale each of the four predictor variables on how important he thought it was in determining testing importance. Figure 3 shows the scale used for rating the four predictor variables on testing importance.

The coefficient of interrater agreement adjusted for differences in the frame of reference for the individual raters was computed as a measure of reliability. The average reliability was found to be .383 for a single rater (R_{11}) and .972 for the means of the 56 raters (R_{kk}). A second sample of 50 raters was later obtained to check the representativeness of the 56-rater sample. No significant difference was found between the adjusted factor weights of the two samples (see Table 2).

The next step was to use the factor rating data to derive an appropriately weighted composite of the four factors that would serve

Figure 3. Scale for Rating Four Factors on Testing Importance

RATING OF FACTOR IMPORTANCE

How important do you think each of the following factors is in determining the testing importance of a task? Use the 9-point rating scale shown below.

The factor is:

1. Extremely unimportant
2. Very unimportant
3. Unimportant
4. Slightly unimportant
5. So-so
6. Slightly important
7. Important
8. Very important
9. Extremely important

Factor	Rating (1-9)
1. % MEMBERS PERFORMING	________
2. LEARNING DIFFICULTY	________
3. CONSEQUENCES OF INADEQUATE PERFORMANCE . .	________
4. REQUIREMENT FOR PROMPTNESS OF PERFORMANCE .	________

DEFINITIONS OF TASK RATING FACTORS

1. % Members Performing is a measure of the proportion of all airmen in the appropriate Air Force Specialty or shredout who perform the task.

2. Learning Difficulty is a measure of the need for lengthy, systematic training before a new member of the appropriate Air Force Specialty or shredout can perform the task adequately. It may be thought of as the difficulty involved in "picking up" the task on the job without any systematic training.

3. Consequences of Inadequate Performance is a measure of the seriousness of the probable consequences of inadequate performance of the task. It is measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours of work, etc.

4. Requirement for Promptness of Performance is a measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it. Must he commence immediately, or does he have time to consult a manual, seek guidance, or even be taught how to do it?

as the criterion of testing importance. However, it was first necessary to standardize each of the factors (mean = 5, S.D. = 1) so that all factors would possess equal weight prior to the application of the rater-derived weights. One additional problem existed in regard to the "percent members performing" factor: it was very negatively skewed in all samples and the standard deviation was approximately equal to the mean. As a result, task percentages below the mean would tend to be underweighted and percentages above the mean overweighted, even after standardization. Therefore, it was necessary to extract the logarithm of this variable prior to standardization in order to reduce the skewness.

A covariance weighting technique was used to adjust the factor weights derived from the ratings of factor importance. This was done so as to insure that the factor weights would be in accord with their relative independence. The procedure used to accomplish the covariance weighting is presented in Appendix A.

Comparison of Card-Sorting and Factor Importance Rating Approaches

The card-sorting policy-capturing approach had two distinct advantages:

1. The ranking of simulated tasks on testing importance involved the simultaneous consideration of all four predictor variables, rather than one at a time.

2. Multiple observations were obtainable on each rater.

On the other hand, the card-sorting policy-capturing technique disclosed several weaknesses:

1. Many subject-matter specialists found the rankordering procedure overly complex and difficult to understand.

2. Many subject-matter specialists were turned off by the fact that they were expected to rankorder sets of four numbers that were not associated with identifiable tasks.

3. Three of the four policies identified through the hierarchical grouping of individual rater policy equations gave 76% to 87% of the testing importance weight to a single factor. Only 15 out of 50 raters used a multiple factor policy. This finding indicated that most raters took the line of least resistance and simply rankordered the simulated tasks on a single variable because that was the easiest thing to do.

4. The rankordering of simulated task decks lacks credibility. The technique is difficult to explain and is, therefore, difficult to justify. The technique also operated something like a black box; you knew what went in and what came out, but were not at all sure what happened in between. By way of overall assessment of the policy-capturing technique, I would not recommend its use with enlisted personnel, and, if used at all, it should be very carefully explained and illustrated with numerous examples. It should also be carefully monitored during the entire time the procedure is being performed.

The factor importance weighting technique had two distinct advantages:

1. Obtaining overall testing importance ratings for the four predictor variables was a relatively quick and simple way of obtaining factor weights from a large number of subject-matter specialists.

2. The statistical techniques used to combine rater weights with covariance data to predict a composite criterion were straightforward and the process was totally visible, which lent it credibility.

The principal weakness of the factor importance weighting technique is that the factor weights specified by the raters may not have represented their desired policy had the raters been able to see the resultant testing importance values. However, the effect of using inadequate weights can be corrected over time by subsequent subject-matter specialists who operationally use the task testing importance values based on the inadequate weights. Suggested changes in the ranking of specific tasks can be translated into weighting revisions in the applicable regression equation.

Comparisons of the beta weights for the card-sorting group and comparable scaled-down mean rating weights for the two factor importance weighting groups are shown in Table 2. Although the two weighting schemes yielded the same rankordering of variables in terms of relative contribution in predicting the criterion, the two weighting systems would produce significantly different R^2 values if applied to the same criterion. Based on the previously stated assessments of the two methods, it would appear that more faith should be placed in weights derived by the factor importance weighting technique, although these weights, too, are suspect until such time as they have been subjected to further validation.

Table 2
Weights Derived From Ranking Technique for One Sample and Rating Techniques for Two Samples

Factor	Card-Sorting Sample N=50	Factor Importance Sample #1 N=56	Factor Importance Sample #2 N=50
% MEMBERS PERFORMING	.1715	.2309	.2684
LEARNING DIFFICULTY	.1650	.2250	.2480
CONSEQ INADEQ PERF	.5622	.3444	.3090
REQ FOR PROMPT PERF*	.1896	.2880	.2629

*Task delay tolerance scale was reversed and called "Requirement for Prompt Performance."

Differences between the ranking weights and the two sets of rater weights are significant beyond .001 level of confidence. Differences between the two sets of factor importance weights are not signficant ($p > .05$).

Development of Task-Based Computerized Outline Formats

Up to this point, I have discussed the selection of the subset of usable tasks, the sorting of these tasks into the three SKT testing levels, and the development of a composite task variable which assigns a testing importance value to each task in the subset. It should be noted that the testing importance value for the same task will vary from one testing level to another because of differences in the percent of members performing the task at each level. Task learning difficulty, however, is by definition, invariant from level to level.

Imposing the structure of a test outline on a set of tasks to be presented in a computerized outline required an important decision as to how the tasks should be organized to form meaningful outline areas. Using the categories or modules from the previous SKT outline seemed to be the obvious solution. However, it soon became apparent that this procedure had four serious drawbacks:

1. The test outlines used to develop previous SKTs are controlled-item documents. To extract the outline areas from these documents for use in a computerized outline would involve serious security problems that would be difficult to control.

2. Outlines for many specialties contain extensive content areas dealing with general principles; e.g., electronic principles, mechanical principles, etc., as well as the more directly job-related categories. Many tasks could not be unambiguously assigned to either the general principles area or the job-related area.

3. Previous test outlines may well be out of date and need considerable revision before reuse. Extensive outline revision would negate one of the primary purposes of developing a computerized outline--saving time.

4. Many test outlines are very personal documents that embody the peculiar characteristics of the team that produced it. As such, it is often unacceptable to a subsequent SKT team.

The second document to be considered was the Air Force Specialty Training Standard (STS), which lists in outline format the various job areas of an Air Force specialty in which OJT is to be conducted. It also specifies the performance and knowledge levels that should be attained in each job area at the apprentice, journeyman, and technician/supervisor skill levels. This document is prepared at the training center responsible for the formal training courses pertaining to the applicable specialty and achieves official Air Force status upon being coordinated and approved at command and Air Staff levels. After initial publication, the STS continues to be updated as needed. The official status of the STS, the manner in which it is prepared and approved, and its currency, seemed to make it an ideal document from which to obtain a framework for organizing tasks into meaningful test outline modules. As good fortune would have it, modular organization of tasks had also become a necessary extension of training priorities research being conducted at the Air Force Human Resources Laboratory. As a result, modular capability was added to the Comprehensive Occupational Data Analysis Programs (CODAP) about six months after its need became critical to the development of the SKT computerized outline. However, in order to group tasks into STS modules, tasks first had to be matched with STS work areas. STS coding of tasks by STS paragraph numbers was to be performed by SKT teams at the same time as they coded tasks for "usability." Directions given to subject-matter specialists for STS coding of tasks are shown in Figure 4. Sample pages from a job inventory booklet illustrating STS coding, usability coding, and write-in tasks are shown in Appendix B. Just as task usability coding was accomplished by team consensus, so also for STS coding. However, more than one STS area was allowed to be assigned to a task if the subject-matter specialists felt that this was necessary. Two job inventory booklets were coded--one booklet was retained for use in SKT outline development and the other was made available to job inventory developers and job analysts at the Occupational Measurement Center. The developers and analysts used the STS and task usability codings, as well as suggested task revisions and additions provided by the SKT team, to assist in developing and updating job inventories, or as an aid in organizing, evaluating, and analyzing job survey data. The job inventory booklet retained for SKT outline development was forwarded for keypunching of the STS and task usability codes. STS work area titles were keypunched at the same time.

Figure 4. Coding Instructions for
Recording STS Areas

CODING INSTRUCTIONS

I. RECORDING STS AREAS

A. For each task, indicate the appropriate STS area, subarea, and sub-subarea in the "Check" column; e.g., 3c(2).

B. If there are no subareas, record only the major area; e.g., 14.

C. If more than one STS area, subarea, or sub-subarea applies, record all of them; e.g., 3c(2), 7c(1, 2, 3, 4).

D. At any level of the STS which does not apply, record a dash; e.g., --- (no STS area applies), 3-- (no subarea or sub-subarea applies), 3c- (no sub-subarea applies).

The Development of Weights for Test Outline Areas

The subset of tasks selected for use on the SKTs was carefully screened for usability; i.e., only tasks with a usability code of "8" or "9" were retained. Assuming that appropriate testing importance weights had also been computed for each task, it seemed logical that the procedure for weighting an outline area was to sum the task testing importance weights for that area and divide this sum by the sum of testing importance weights for all tasks in the SKT subset. This computation was carried out in each outline area to transform the sums into proportionate weights, and the proportionate weights were easily convertible into quotas for the number of items to be written per outline area (total of 135 per SKT, 80 per Apprentice Knowledge Test) and the number of items that would actually be selected for use in the test per outline area (total of 100 per SKT, 65 per AKT). For tasks that had been coded into more than one STS area, the procedure was to divide the testing importance weight for each task by the number of assigned STS areas and use this partial weight in the summing operation. As of now, only the summing of the testing importance weights by outline area can be accomplished by computer.

The computation of the proportionate weights and the conversion of proportionate weights to numbers of items to write and to select must be done by hand. Hopefully, the necessary computer programming to replace the manual operation will be accomplished within the next several months.

Actualization of the Computer-Derived Outline

All the components required for the computer-derived outline were now ready for assembly into a printed output. Two types of format were decided upon: one for use with a vertical method of test construction, and another for use with a horizontal method. In the vertical method, SKTs are constructed for one testing level at a time. A print format was designed for this method that consisted of three separate outlines, each of which listed only the tasks to be tested at that level in order of testing importance. Appendix C shows a portion of a 5-level (E-5) outline for use with the vertical method of test construction. In the horizontal method, all levels of SKTs are constructed simultaneously, with test items being assigned to the appropriate level as they are being written. For this method, the most appropriate outline was one which presented the outlines for all levels side by side in a single document. The entire set of tasks to cover all levels was listed, with zero testing importance values showing for a task at the level(s) for which the task was not appropriate. Tasks were listed within each STS module in job inventory sequence, rather than ordered on testing importance. Large open spaces were provided to the right of each column of testing importance values to allow room for test item numbers or other information to be recorded. Appendix D shows a portion of a combined 5-level (E-5) and 7-level (E-6/E-7) outline designed for use with the horizontal method of test construction.

The printed output for either format is the product of the CODAP MODCHK and FACPRT programs. An "executive summary" option has recently been added to FACPRT, which performs the summation of the testing importance values by STS module. In the near future, the executive summary option will also be able to compute and display quotas for the number of test items to be written and the number of test items to be selected for each STS module. Appendix E shows an example of an executive summary as it is expected to appear.

In a paper which immediately follows this one, Capt Conrad Bills will assess the usability of the computer-derived test outlines, based on his experiences in using them in several SKT test construction projects.

Concluding Discussion

Perhaps the weakest link in the development of a test outline from occupational survey data is that the task data do not specify the kind or degree of knowledge required to successfully perform each task. Nevertheless, the computerized outlining procedure presented in this paper can be justified in several ways. First of all, the use of task-based data can be justified on the grounds that the Equal Employment Opportunity Coordinating Council (EEOCC) guidelines are better served by a test based directly on task data. Such a test is ostensibly more job-related than a test based on knowledge requirements, because knowledge requirements are at least one step removed from the task level and are more subjective. Secondly, the task statements in an outline can be viewed as stimuli and the task data as guidelines in directing subject-matter specialists toward selecting and emphasizing in their test item writing those knowledges that are most pertinent to the job. In this model, subject-matter specialists are viewed as the link between task specifications, as laid out in the task-based outline, and knowledge specifications, as determined by work experience and reference materials. Thirdly, the computerized outlining procedure can be defended as a generalized procedure that is able to incorporate job knowledge requirements in the outlining process with little difficulty when such information is available. In the electronics career fields, for example, the electronics principles inventory developed by O'Connor, Ruck, and Driskill (1975) could be interfaced with specially screened task lists in such a way as to attach an electronics theory section onto task-based outlines for critical tasks. The resultant outline would, in fact, more closely resemble the conventional outlines developed by subject-matter specialists, who typically include a theory section. The Plan of Instruction (POI) generated at the training center for each formal course could also be coded to the task list to provide a detailed interface between tasks and knowledge requirements.

Another problem relating to knowledge requirements is that of tasks which have overlapping knowledge requirements. Why, for example, should two tasks with heavily overlapping knowledge requirements be allowed to have separate testing importance weights and thereby make independent contributions to the computation of testing importance weights for STS modules? It could just as easily be asked, "Why not?" If an item of knowledge is applicable to two important tasks, its importance is better reflected by allowing it the summed weight of both tasks than by limiting its weight to that of one task. One exception to this rule would be redundant tasks. The number of redundant tasks, however, will be few in a well-constructed job inventory. What redundancy exists should be virtually eliminated by the task usability coding process. Code "4" is intended to filter out redundant tasks. Even if knowledge requirements were available for tasks, it would be virtually impossible to determine the degree of knowledge overlap between two tasks with similar knowledge requirements.

Another major area of concern has been the job survey data itself. Subject-matter specialists have frequently complained that the inventory data are incomplete, outdated, or inaccurate. In the case of surveys more than two years' old, outdatedness can be a serious problem. However, unless there are evidences of extensive career field changes, it is more likely than not that survey data based on the responses of hundreds, and perhaps thousands of job incumbents, is still more accurate overall than the limited experience of several subject-matter specialists. Incompleteness of the inventory task list is an area that can best be handled by adding and weighting the missing tasks, which are requested as part of the task usability coding process (see sections B and C of Figure 1). As stated previously, the write-in tasks are forwarded to the job inventory developers to be considered for inclusion in the next task inventory. This interplay between the testing process and the inventory development process should, in time, accrue to the benefit of both.

The validity of the weighted composite as a measure of testing importance remains an area of continuous evaluation. While the weights derived from the original sample of 56 raters had high interrater agreement (.972), there is no guarantee that the sample was representative. A subsequent sample of 50 raters produced weights that were not significantly different from the weights derived from the 56-rater sample (see Table 2). One possible solution to the weighting problem would be to gather factor ratings from one complete year of SKT test development teams (four personnel from each of approximately 250 specialties and shredouts). Not only would the sample be large and representative, but another look could be taken at the possibility of finding differential rating policies attributable to specialty, career field, command, grade, or other variables. Differential rating policies that can be tied to specific variables can be translated into weighted testing importance composites tailored to the specific outline requirements of each SKT. Such a project would not be too costly in time or manpower. Obtaining the factor ratings would take about ten minutes per SKT team, including instructions, and would ideally take place at the conclusion of a test development project. During the year of data gathering, the weighted equation currently in use could continue to be updated as additional ratings are obtained.

The current testing importance equation is confined to "percent of members performing" and "task learning difficulty" for most specialties because of the frequent unavailability of data on "probable consequences of inadequate performance" and "task delay tolerance." On the other hand, there is a continual buildup of data on the "recommended training emphasis" variable. While it does not appear that training emphasis can be used as a substitute for testing importance, as discussed earlier in this paper, studies should be made to compare training emphasis with testing importance to determine what generalizations can be made concerning similarities and differences. It may well be that training emphasis could play an important role in improving estimates of testing importance of tasks.

Another important area in which potential problems exist is that of task selection criteria. The need to use the major command variable in selecting tasks came to light only after it was discovered that a whole block of command-specific tasks had been added to a computerized outline because no selection criterion had been applied which required that incumbents performing a task be representative of all the major using commands. As experience with the computerized outline grows, other necessary selection criteria will undoubtedly come to light, and current selection criteria will have to be augmented. Some of the new criteria may be general, others may be specific to a specialty or career field.

Although the computerized test development outline is intended to be a stand-alone product, it is not intended to be the only computer product used. It would be foolish for an SKT team not to use other available occupational survey data, such as the variable summary (VARSUM) which contains information on tools, equipment, manuals, and procedures used by job incumbents, as well as other information pertinent to the test construction process.

Various time-saving devices are under consideration to make the computerized outlining technique more cost effective. One such device is to provide the subject-matter specialists who perform the STS and task usability coding with a task list from which low performance, supervisory, and command-specific tasks have been eliminated, in lieu of the current requirement that all tasks in the job inventory booklet be coded. This would probably cut the normal four-hour coding time to no more than an hour. Another time-saving device would be to simplify and automate as much of the computer runstream as possible. Currently, as many as ten separate computer runs are being made to produce the final outline product. A third time and cost saver would be to develop standardized, self-explanatory forms for requesting a computerized outline. This would reduce request preparation time and would permit the use of low-pay clerical personnel to prepare the requests.

The ultimate computerized outline document is at least several months away. This document will not only provide the listing of tasks

1014

within STS modules and the testing importance values for tasks, but also the number of test items to write and select on each task. In addition, a summary report will be provided that will list the STS modules in outline format and the percentage weight computed for each module, along with quotas for the number of test items to be written and the number of test items to be selected for each module.

Since the development and implementation of the new test outline technique described in this paper is an ongoing, incremental, and interactive process, occasional modifications will be required, and a few specialties may present insurmountable difficulties. Nevertheless, the procedure is viable, and the alternative--failure to make adequate use of occupational data in test development--is no longer acceptable.

1015

References

Bills, Conrad G. Evaluation of computer-derived test outlines using conventional test outlines as a criterion reference during test development projects. Paper presented to 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 October-3 November 1978.

Christal, R.E. JAN: A technique for analyzing group judgment. *Journal of Experimental Education*, 1968, 36(4), 24-27.

Christal, R.E. *The United States Air Force occupational research project*. AFHRL-TR-73-75, AD-774 574. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Christal, R.E. & Weissmuller, J.J. *New CODAP programs for analyzing task factor information*. AFHRL-TR-76-3, AD-A026 121. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, May 1976.

DuBois, Philip H. *An Introduction to Psychological Statistics*. New York: Harper & Row, 1965, 215-221.

Equal Employment Opportunity Coordinating Council. *Uniform Guidelines on Employment Selection Procedures*, June 1974.

Ghiselli, Edwin E. *Theory of Psychological Measurement*. New York: McGraw-Hill Book Company, Inc., 1964, 187-192.

Gott, C.D. *HIER-GRP: a computer program for the hierarchical grouping of regression equations*. AFHRL-TR-78-14. Brooks Air Force Base, TX: Computational Sciences Division, Air Force Human Resources Laboratory, June 1978.

Guilford, J.P. & Fruchter, Benjamin. *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill Book Company, Inc., 1973, 261-264 and 381-382.

Haggard, E.A. *Intraclass Correlation and the Analyses of Variance*. New York: Dryden Press Inc., 1958, 12-15.

Lindquist, E.F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953, 359-361.

Morsh, J.E. & Christal, R.E. *Impact of the computer on job analysis in the United States Air Force*. PRL-TR-66-19, AD-656 304. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, October 1966.

O'Connor, Thomas J., Ruck, Hendrick W. & Driskill, Walter E. A universal model for evaluating basic electronic courses in terms of field utilization of training. Paper presented to 17th Annual Conference of the Military Testing Association, U.S. Army, Indianapolis, IN, 16-19 September 1975.

Peters, Charles C. & Van Voorhis, Walter R. Statistical Procedures and Their Mathematical Bases. New York: McGraw-Hill Book Company, Inc., 1940, 220-245.

Ruck, H.W., Thompson, N.A. & Thomson, D.C. The collection and prediction of training emphasis ratings for curriculum development. Paper presented to 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 October-3 November 1978.

Stacey, William D., Weissmuller, Johnny J., Barton, Bruce B. & Rogers, C.R. CODAP: Control Card Specifications for the UNIVAC 1108. AFHRL-TR-74-84. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, October 1974.

Stacy, W.J., Thompson, N.A. & Thomson, D.C. Occupational task factors for instructional systems development. Paper presented to 19th Annual Conference of the Military Testing Association, USAF, San Antonio, TX, 17-21 October 1977.

Thew, M.C. CODAP: A modular approach to occupational analysis. Paper presented to the 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 October-3 November 1978.

USAF Occupational Measurement Center. Handbook for Construction of the SKT and Associated Tests. Lackland AFB, TX: author, June 1, 1977.

USAF Occupational Measurement Center. Subject-Matter Specialist Handbook for Construction of the SKT and Associated Tests. Lackland AFB, TX; author, July 1, 1977.

Vaughan, D.S. Prediction of Test Outline Weights from Occupational Survey Data. Proceedings of the 18th Annual Conference of the Military Testing Association, 1976, 435-460.

Vaughan, D.S. The Interface Between Occupational Survey and Test Construction. Proceedings of the 19th Annual Conference of the Military Testing Association, 1977, 798-800.

Ward, J.H. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244.

Appendix A
Factor Covariance Weighting Technique

With factor importance weights computed and factor data standardized, the next step in the factor covariance weighting technique was to determine the correlation of each weighted factor with the four-variable composite criterion, including in the computation the known covariances of the component variables. This was accomplished by applying the following equation to each component variable of the composite, using as input the four-variable variance-covariance matrix for a specific Air Force specialty:

$$\frac{r_{1t} = w_1{}^2 + \sum_{i=1}^{n-1} w_1 w_2 r_{1j}}{w_1 \sqrt{\sum_{i=1}^{n} w_i{}^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j r_{ij}}} \qquad i \neq j$$

where

r_{1t} = correlation of variable 1 with composite

w_1 = rater-derived weight for variable 1

$\sum_{i=1}^{n-1} w_1 w_j r_{1j}$ = the sum of the cross-products of rater-derived weights and covariances which involve variable 1

$\sum_{i=1}^{n} w_i{}^2$ = the sum of the squared weights for the "n" variables

$2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j r_{ij}$ = the sum of the cross-products of weights and covariances for the entire variance-covariance matrix

NOTE: Since the variance elements in the variance-covariance matrix = 1.00, they have been dropped from all cross-products in the equation.

With the correlations between each variable and the composite criterion computed, it was then possible to regress the four component variables on the composite to arrive at their appropriate standard score weights. These weights would, of course, vary from specialty to specialty. When actually computing testing importance values for tasks, the criterion parameters were standardized to mean = 5 and S.D. = 2.[3] These parameters were chosen to satisfy several requirements:

1. To standardize and simplify the interpretation of task testing importance values across all specialties.

2. To identify tasks of very low testing importance which would later be eliminated from inclusion in the computerized test outline. This was accomplished by setting the testing importance of a task equal to zero if the computed testing importance was less than zero (more than -2.50 S.D. below the mean).

3. To maximize the variance of the criterion composite without deviating from requirements 1 and 2. Maximizing the variance not only added visual emphasis to differences in testing importance between tasks, but also reduced the mean (relative to the variance) in the calculation of outline area weights. Weighting of the test outline will be discussed later.

In actual practice, "probable consequences of inadequate performance" and "task delay tolerance" data were not available for the specialties in which there was an opportunity to experiment with this second technique. As a result, the computation of task testing importance included only the "percent members performing" and "task difficulty" variables. Even here, the factor covariance weighting technique was applicable and differences in the single covariance value produced differences in the standard score weights for the two-variable composite.

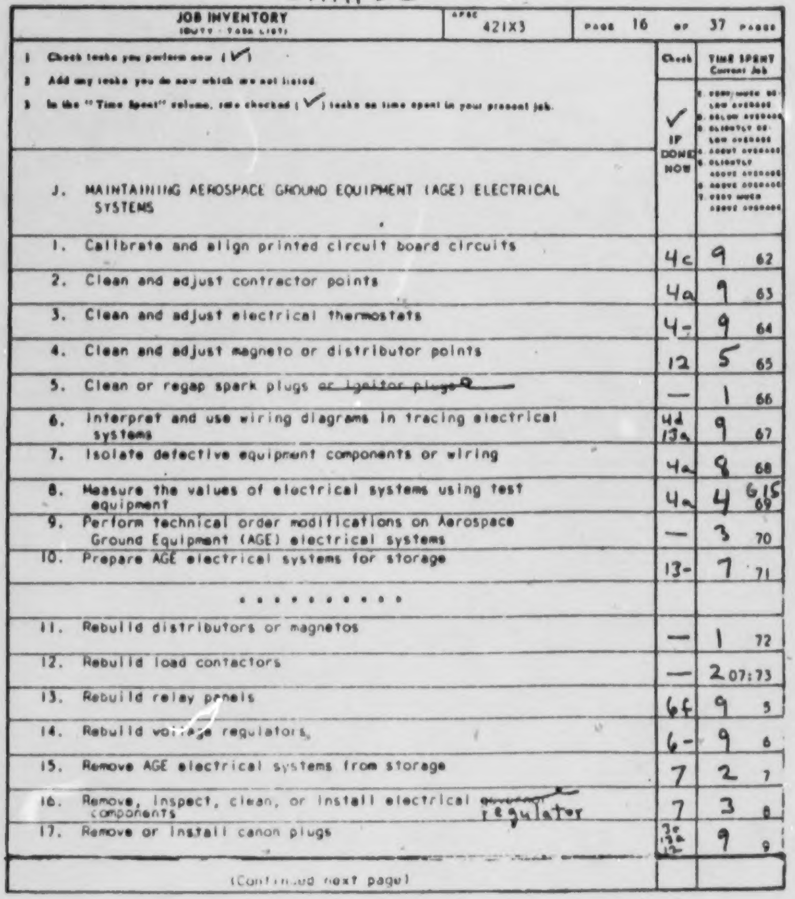
---

[3] Since each of the four factors in the composite were standardized to mean = 5 and S.D. = 1, the mean of the composite would also equal 5. To set the standard deviation of the composite equal to 2, all that was necessary was to multiply each of the four beta weights by 2 before computing the composite.

Appendix B

Sample Pages from Job Inventory Booklet Illustrating STS Coding, Usability Coding, and Write-in Tasks

SAMPLE

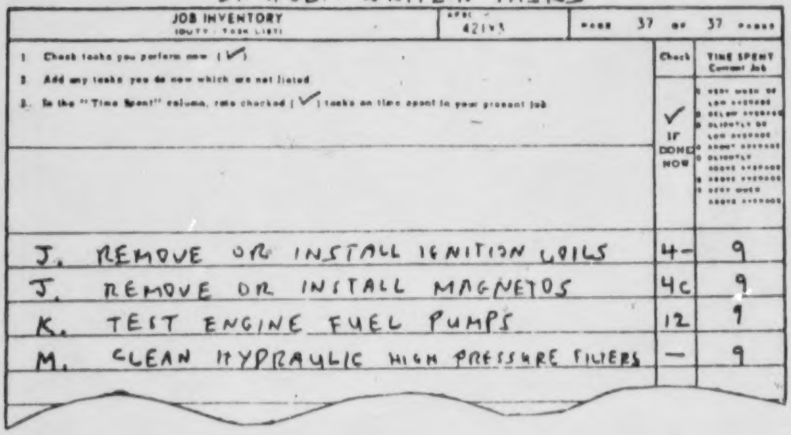


JOB INVENTORY (DUTY - TASK LIST)	AFSC 421X3	PAGE 16 OF 37 PAGES

1 Check tasks you perform now (✓)
2 Add any tasks you do now which are not listed
3 In the "Time Spent" column, rate checked (✓) tasks on time spent in your present job.

	Check ✓ IF DONE NOW	TIME SPENT Current Job 1. Very much below average 2. Below average 3. Slightly below average 4. About average 5. Slightly above average 6. Above average 7. Very much above average	
J. MAINTAINING AEROSPACE GROUND EQUIPMENT (AGE) ELECTRICAL SYSTEMS			
1. Calibrate and align printed circuit board circuits	4c	9	62
2. Clean and adjust contractor points	4a	9	63
3. Clean and adjust electrical thermostats	4-	9	64
4. Clean and adjust magneto or distributor points	12	5	65
5. Clean or regap spark plugs ~~or ignitor plugs~~	—	1	66
6. Interpret and use wiring diagrams in tracing electrical systems	4d 13a	9	67
7. Isolate defective equipment components or wiring	4a	8	68
8. Measure the values of electrical systems using test equipment	4a	4	6.15 69
9. Perform technical order modifications on Aerospace Ground Equipment (AGE) electrical systems	—	3	70
10. Prepare AGE electrical systems for storage	13-	7	71
• • • • • • • • • •			
11. Rebuild distributors or magnetos	—	1	72
12. Rebuild load contactors	—	2	07:73
13. Rebuild relay panels	6f	9	5
14. Rebuild voltage regulators	6-	9	6
15. Remove AGE electrical systems from storage	7	2	7
16. Remove, inspect, clean, or install electrical ~~governor~~ regulator components	7	3	8
17. Remove or install canon plugs	3c 13a 12-	9	9
(Continued next page)			

SAMPLE - WRITEIN TASKS



JOB INVENTORY (DUTY - TASK LIST)	AFSC 421X3	PAGE 37 OF 37 PAGES

1 Check tasks you perform now (✓)
2 Add any tasks you do now which are not listed
3 In the "Time Spent" column, rate checked (✓) tasks on time spent in your present job

	Check ✓ IF DONE NOW	TIME SPENT Current Job 1. Very much below average 2. Below average 3. Slightly below average 4. About average 5. Slightly above average 6. Above average 7. Very much above average
J. REMOVE OR INSTALL IGNITION COILS	4-	9
J. REMOVE OR INSTALL MAGNETOS	4c	9
K. TEST ENGINE FUEL PUMPS	12	9
M. CLEAN HYDRAULIC HIGH PRESSURE FILTERS	—	9

Appendix C
Computerized Outline of E-6/E-7 Testing Importance for Use with the Vertical Method of Test Construction

E-6/E-7 TESTING IMPORTANCE - ORDERED HIGH TO LOW (AFSC 701X0) FCPRT3 PAGE 7

AF HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND

D	TSK	TITLES	TST IM7	TSK DIF	PCT PRF
035	6B1	PREPARE AND RESTORE RELIGIOUS FACILITIES, EQUIPMENT, AND APPOINTMENTS FOR RELIGIOUS SERVICES			
H	19	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC WORSHIP SERVICES	5.81	4.23	54.9
H	17	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC SACRAMENTAL RITES	5.69	4.28	50.0
H	15	PREPARE CHAPEL FACILITIES FOR ENCUMENICAL SERVICES	5.44	4.11	48.6
H	25	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT WORKSHIP SERVICES	5.43	3.95	53.5
H	23	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT SACRAMENTAL RITES	5.37	4.06	48.6
I	15	RESTORE CHAPEL FACILITIES AFTER USE	4.85	3.43	54.2
H	26	PREPARE CHAPEL FACILITIES TO SUPPORT INTERFAITH SERVICES/ACTIVITIES	4.63	4.13	31.9
H	20	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SACRAMENTAL RITES	4.62	4.27	29.2
H	[illegible]	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SERVICES	4.52	4.32	27.1
H	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	4.51	4.18	29.2
H	9	ORGANIZE LAY PERSONNEL TO SUPPORT SACRAMENTAL RITES	4.51	4.80	20.1
H	8	NEUTRALIZE CHAPEL ALTAR AFTER SERVICES	4.38	3.19	49.3
I	4	CLEAN ECCLESIASTICAL EQUIPMENT	3.76	3.12	37.5
036	6B2	---RELIGIOUS EDUCATION			
H	2	COORDINATE WITH LAY PERSONNEL IN SUPPORT OF RELIGIOUS EDUCATION ACTIVITIES	6.38	4.56	59.7
I	15	RESTORE CHAPEL FACILITIES AFTER USE	4.85	3.43	54.2
H	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	4.51	4.18	29.2
H	32	PREPARE FACILITIES FOR RELIGIOUS EDUCATION ACTIVITIES SUCH AS PRE-MARRIAGE OR PARENT EFFECTIVENESS TRAINING (P.E.T.).	4.21	3.85	30.6
H	4	MAINTAIN RELIGIOUS EDUCATION CURRICULUM CATALOGS	3.60	3.62	25.7
H	3	ISSUE RELIGIOUS EDUCATION MATERIALS OR SUPPLIES	3.59	3.79	22.9
037	6B3	---RELATED CHAPEL ACTIVITIES			
H	16	PREPARE CHAPEL FACILITIES FOR MEMORIAL/FUNERAL SERVICES	5.77	4.32	50.7
I	15	RESTORE CHAPEL FACILITIES AFTER USE	4.85	3.43	54.2
H	18	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC SPIRITUAL RENEWAL PROGRAMS	4.82	4.29	31.9
H	24	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT SPIRITUAL RENEWAL PROGRAMS	4.65	4.19	31.3
H	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	4.51	4.18	29.2
H	30	PREPARE FACILITIES AND EQUIPMENT FOR ADULT VALUE EDUCATION	4.15	4.13	25.0

Appendix D

Computerized Outline of Two Levels of Testing Importance for Use with Horizontal Method of Test Construction (Two Pages)

```
TWO LEVELS OF TEST IMPORTANCE IN STS ORDER                          FCPRT3 PAGE     2        AF HUMAN RESOURCES LABORATORY
                                                                                               AIR FORCE SYSTEMS COMMAND
PERCENT MEMBERS PERFORMING TASKS MATCHED TO STS 701X0 PARAGRAPHS
INVENTORY RESTRICTED TO TASKS FOR WHICH THE
      MAXIMUM OF THE FIRST  2 INPUT VECTORS IS GT .000

VECTOR TYPE CODES:

(T) = % TIME SPENT BY ALL MEMBERS
(M) = % MEMBERS PERFORMING
(F) = TASK FACTOR
(D) = DICHOTOMOUS SET
(R) = % TIME SPENT BY MEMBERS PERFORMING
(-) = PROGRAM GENERATED VECTOR

COL  TYPE  VECTOR   MEMRS/MEAN  DESCRIPTION                                          FACTOR #

  1    F   TSTIM5         1.69  E-5 STANDARDIZED TESTING IMPORTANCE VALUES              195
  2    F   TSTIM7         1.39  E-6/E-7 STANDARDIZED TESTING IMPORTANCE VALUES          196
  3    -   MAXIMA               MAXIMUM VALUE OF THE FIRST  2 INPUT VECTORS
  4    F   TSKFC2         7.79  TEST-USE WEIGHT DECK                                    193
  5    F   TSKFC1         5.00  TASK DIFFICULTY INTERRATER RELIABILITY                  121
  6    M   SPL022   109         ALL AMN 701X0 IN PAYGRADE E5                             27
  7    M   SPL025   144         ALL AMN 701X0 IN PAYGRADE E6 AND E7 COMBINED             30
```

```
TWO LEVELS OF TEST IMPORTANCE IN STS ORDER                          FCPRT3 PAGE     3        AF HUMAN RESOURCES LABORATORY
                                                                                               AIR FORCE SYSTEMS COMMAND
PERCENT MEMBERS PERFORMING TASKS MATCHED TO STS 701X0 PARAGRAPHS
INVENTORY RESTRICTED TO TASKS FOR WHICH THE
      MAXIMUM OF THE FIRST  2 INPUT VECTORS IS GT .000

                                                          TST   ITEMS  ITEMS   TST   ITEMS  ITEMS   TSK   SPL   SPL
                                                          IM5    TO     TO     IM7    TO     TO     FC1   022   025
                                                          (F)   WRITE SELECT   (F)   WRITE SELECT   (F)   (M)   (M)
-----------------------------------------------------------------------------------------------
U34    6A  PLAN, ORGANIZE, DIRECT, AND CONTROL SUPPORT OF CHAPEL
              PROGRAM                                               5      4            7      5
-----------------------------------------------------------------------------------------------
M     1 ASSIST IN PREPARATION OF REFRESHMENTS FOR CHAPEL SOCIAL  4.67                4.47                 3.44  69.7  44.4
         FUNCTIONS
M    33 PROVIDE FINANCIAL, MATERIEL, OR ADMINISTRATIVE SUPPORT   6.96                6.66                 5.13  74.3  50.0
         TO LAY PERSONNEL
```

TWO LEVELS OF TEST IMPORTANCE IN STS ORDER

FCPNT3 PAGE 7

AF HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND

D	TSK	TITLES	TST IMP (F)	ITEMS TO WRITE	ITEMS TO SELECT	TST IMP (F)	ITEMS TO WRITE	ITEMS TO SELECT	TSK FCI (F)	SPL 022 (M)	SPL 025 (M)
D35	681	PREPARE AND RESTORE RELIGIOUS FACILITIES, EQUIPMENT, AND APPOINTMENTS FOR RELIGIOUS SERVICES		35	26		38	28			
M	8	NEUTRALIZE CHAPEL ALTAR AFTER SERVICES	4.67			4.38			3.19	80.7	49.3
M	9	ORGANIZE LAY PERSONNEL TO SUPPORT SACRAMENTAL RITES	4.66			4.51			4.80	31.2	20.1
M	15	PREPARE CHAPEL FACILITIES FOR ECUMENICAL SERVICES	5.80			5.44			4.11	78.9	48.6
M	17	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC SACRAMENTAL RITES	6.10			5.69			4.20	81.7	50.0
M	19	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC WORSHIP SERVICES	6.03			5.81			4.23	81.7	54.9
M	20	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SACRAMENTAL RITES	5.11			4.62			4.27	52.3	29.2
M	21	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SERVICES	5.10			4.52			4.32	50.5	27.1
M	23	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT SACRAMENTAL RITES	5.79			5.37			4.06	80.7	48.6
M	25	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT WORSHIP SERVICES	5.70			5.43			3.95	82.6	53.5
M	26	PREPARE CHAPEL FACILITIES TO SUPPORT INTERFAITH SERVICES/ACTIVITIES	5.19			4.63			4.13	58.7	31.9
M	27	PREPARE CHAPEL FACILITIES TO SUPPORT JEWISH SACRAMENTAL RITES	3.82			.00			4.55	29.8	9.0
M	29	PREPARE CHAPEL FACILITIES TO SUPPORT JEWISH WORSHIP SERVICES	4.09			.00			4.41	30.3	15.3
M	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	5.25			4.51			4.16	58.7	29.2
I	4	CLEAN ECCLESIASTICAL EQUIPMENT	4.29			3.76			3.12	70.6	37.5
I	15	RESTORE CHAPEL FACILITIES AFTER USE	5.11			4.85			3.43	85.3	54.2
D36	682	---RELIGIOUS EDUCATION		15	11		16	12			
M	2	COORDINATE WITH LAY PERSONNEL IN SUPPORT OF RELIGIOUS EDUCATION ACTIVITIES	5.08			6.38			4.56	62.4	59.7
M	3	ISSUE RELIGIOUS EDUCATION MATERIALS OR SUPPLIES	4.41			3.59			3.79	50.5	22.9
M	4	MAINTAIN RELIGIOUS EDUCATION CURRICULUM CATALOGS	3.80			3.60			3.62	42.2	25.7
M	31	PREPARE FACILITIES FOR RELIGIOUS EDUCATION ACTIVITIES SUCH AS TRANSACTIONAL ANALYSIS (T.A.)	3.40			.00			3.84	37.6	19.4
M	32	PREPARE FACILITIES FOR RELIGIOUS EDUCATION ACTIVITIES SUCH AS PRE-MARRIAGE OR PARENT EFFECTIVENESS TRAINING (P.E.T.)	4.27			4.21			3.85	45.9	30.6
M	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	5.25			4.51			4.16	58.7	29.2
I	15	RESTORE CHAPEL FACILITIES AFTER USE	5.11			4.85			3.43	85.3	54.2
D37	683	---RELATED CHAPEL ACTIVITIES		14	10		17	13			
M	16	PREPARE CHAPEL FACILITIES FOR MEMORIAL/FUNERAL SERVICES	5.87			5.77			4.32	71.6	50.7
M	18	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC SPIRITUAL RENEWAL PROGRAMS	5.21			4.82			4.29	54.1	31.9
M	24	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT SPIRITUAL RENEWAL PROGRAMS	5.15			4.65			4.14	56.0	31.3
M	30	PREPARE FACILITIES AND EQUIPMENT FOR ADULT VALUE EDUCATION	4.85			4.15			4.13	50.5	25.0
M	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	5.25			4.51			4.18	58.7	29.2
I	15	RESTORE CHAPEL FACILITIES AFTER USE	5.11			4.85			3.43	85.3	54.2

Appendix D (Page Two)

Appendix E
Executive Summary for Two Levels of Testing Importance
Ordered by STS Paragraph Number

TWO LEVELS OF TEST IMPORTANCE IN STS ORDER - EXECUTIVE SUMMARY - AFSC 701X0

STS NUM	TITLE OF STS MODULE	NUM OF TSK	SUM TST IM5	NUM ITM TO WRT	NUM ITM TO SLT	NUM OF TSK	SUM TST IM7	NUM ITM TO WRT	NUM ITM TO SLT
3A6	MAINTAIN RECORDS AND GRAPHS TO SUPPORT REQUIREMENTS	1	6.70	2	2	1	6.41	2	2
3A10	PREPARE, PROCESS, AND MAINTAIN APPROPRIATED FUND REQUESTS, CONTRACTS, AND RECORDS	5	30.40	9	7	5	30.56	11	8
3B1	PLAN, ORGANIZE, DIRECT, AND CONTROL FACILITY UTILIZATION	1	4.71	1	1	1	5.38	2	2
3B7	INSPECT FACILITIES, INCLUDING FIRE, HEALTH, AND SAFETY HAZARDS, AND TAKE CORRECTIVE ACTION	1	5.24	2	1	1	5.62	2	2
3C4	EXERCISE SUPPLY DISCIPLINE, PROPERTY ACCOUNTABILITY AND RESPONSIBILITY	2	11.25	3	3	2	11.73	4	3
5D2	OPERATE 16MM PROJECTOR	1	6.26	2	1	1	6.21	2	2
5D3	OPERATE 35MM FILM STRIP AND SLIDE PROJECTOR	1	6.26	2	1	1	6.21	2	2
5D5	OPERATE PUBLIC ADDRESS SYSTEMS	1	4.86	1	1	1	5.24	2	1
6A	PLAN, ORGANIZE, DIRECT, AND CONTROL SUPPORT OF CHAPEL PROGRAM	2	11.65	3	3	2	11.15	4	3
6B1	PREPARE AND RESTORE RELIGIOUS FACILITIES, EQUIPMENT, AND APPOINTMENTS FOR RELIGIOUS SERVICES	15	76.71	23	17	13	63.52	23	17
6B2	---RELIGIOUS EDUCATION	7	32.62	10	7	6	27.14	10	7
6B3	---RELATED CHAPEL ACTIVITIES	6	31.44	9	7	6	28.75	10	8
6C1	PROVIDE GUIDANCE AND TRAINING TO LAY PERSONNEL IN SUPPORT OF RELIGIOUS SERVICES	1	5.11	2	1	1	4.85	2	1
6C2	---RELIGIOUS EDUCATION PROGRAMS	2	5.11	2	1	2	4.85	2	1
6C3	---SPIRITUAL RENEWAL ACTIVITIES	2	9.87	3	2	2	9.77	3	3
6C4	---STEWARDSHIP AND HUMANITARIAN PROJECTS	1	9.69	3	2	1	9.75	3	3
6C5	---SOCIAL CONCERN ACTIVITIES	1	5.11	2	1	1	4.85	2	1
6C6	---RELATED CHAPEL ACTIVITES	1	5.11	2	1	1	4.85	2	1
7A1	DRAFT, TYPE, AND PROCESS LETTTERS	1	4.59	1	1	1	5.23	2	1
7A2	---MESSAGES	5	4.59	1	1	5	5.23	2	1
7A3	---FORMS	3	26.44	8	6	3	26.53	10	7
7A4	---PUBLICITY MATERIAL	1	14.69	4	3	1	15.49	6	4
7A5	---REPORTS	1	4.59	1	1	1	5.23	2	1
7A6	---ORDERS	1	4.59	1	1	1	5.23	2	1
7B2	MAINTAIN FILES	2	9.19	3	2	2	9.32	3	3
7C3	DETERMINE AND ESTABLISH FORMS REQUIREMENTS	1	5.22	2	1	1	5.04	2	1
8A1A	RECORD, TYPE, AND PROCESS AGENDA AND FUND COUNCIL MINUTES	2	10.48	3	2	0	0.00	0	0
8A1B	---PURCHASE ORDERS	1	5.00	1	1	0	0.00	0	0
8A1E	---RELATED FINANCIAL RECORDS	3	15.21	5	4	1	7.28	3	2
8A1F	---RECEIPTS	3	12.18	4	3	1	3.19	1	1
8A2	ESTABLISH, MONITOR, AND MAINTAIN PROCEDURES TO SAFEGUARD MONEY AND MATERIAL	4	20.17	6	5	2	9.24	3	3
8A3	MAINTAIN FUND ACCOUNTING SYSTEM	1	6.52	2	2	0	0.00	0	0
8A6	PREPARE, TYPE, AND PROCESS CONTRACTS	1	5.38	2	1	0	0.00	0	0
8B3	DEVELOP, IMPLEMENT, AND MAINTAIN PROCEDURES TO SAFEGUARD PROPERTY	1	4.76	1	1	1	4.24	1	1
---	TASKS NOT REFERENCED	8	28.70	9	6	7	26.50	10	7
	TOTAL	90	450.40	135	100	75	374.57	135	100

Evaluation Of Computer-Derived Test Outlines Using Conventional Test Outlines As a Criterion Reference During Test Development Projects

Conrad G. Bills

USAF Occupational Measurement Center
Lackland Air Force Base, Texas

A paper presented at the 20th Annual Conference of the
Military Test Association
November 1978

1027

EVALUATION OF COMPUTER-DERIVED TEST OUTLINES USING CONVENTIONAL TEST OUTLINES AS A CRITERION REFERENCE DURING TEST DEVELOPMENT PROJECTS[1]

Conrad G. Bills, Capt, USAF
USAF Occupational Measurement Center

At the 1976 and the 1977 Military Testing Association conferences, Vaughan (1976, 1977) described the interrelationship between test construction and occupational surveying activities at the USAF Occupational Measurement Center. As part of the cross-feed between these two activities, Vaughan (1977) described a procedure for the automated conversion of occupational survey data into a test outline. This computer-derived outline would indicate the number of test items to be written on each topic. There were two procedures that had been attempted and he mentioned that a synthesis of these procedures was being tested. William J. Phalen (1978) has described the development of this synthesized technique for using occupational survey data to construct and weight computer-derived test outlines. This technique is designed to increase the relative ease with which occupational survey data can be incorporated into the test construction process. The incorporation of survey data will in turn strengthen the content validity position (Vaughan, 1977) of these tests which are under the Weighted Airman Promotion System (WAPS). Under the proposed EEOC guidelines (1977), the need for a strong validity position is paramount. The purpose of this study was to evaluate the experimental application of the computer-derived test outline procedure using the conventional outline as a criterion reference.

Conventional Outline

The conventional outline development procedure described by Vaughan (1976) has been used consistently over two decades (USAF Occupational Measurement Center, 1977). An average test development team consists of four subject-matter specialists (SMSs). SMSs are first asked to divide their job specialty into major divisions. These divisions constitute the major outline areas. The major outline areas are then subdivided as appropriate. Once the team members have reached agreement, they are asked to assign percentage weights to each outline area. The resultant percentage weights determine the number of test questions to be written for each division of the job specialty. A sample of the outline format is shown in Figure 1. Percentage weights are determined by SMSs, based

Insert Figure 1 about here

on their knowledge and experience. Their judgment is supplemented by the occupational survey data provided to them. Since test construction teams have found survey data difficult to use, the contribution of survey data to the outline development process has been minimal (Vaughan, 1976).

---

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

Computer-Derived Outline

The procedure for developing a computer-derived test outline differs from the procedure for developing the conventional outline. The initial computer printout displays selected occupational survey tasks with testing importance values. A sample page from a computer printout is shown in Figure 2. These tasks are presorted by Specialty Training Standard (STS) paragraph. Therefore, the test construction team only evaluates the printout and finalizes the major outline areas. The SMSs can adjust the task sorting for major outline areas. They can also adjust the percentage weights that have been determined by the testing importance values. A conversion table is given to the team for determining the equivalent percentage weights from the testing importance values that are on an initial computer product (Table 1). The team must justify the changes they make to the computer product. Like the conventional outline, the resultant percentage weights determine the number of test questions to be written for each division of the job-specialty. Unlike the conventional outline, occupational survey data is the basis for computer-derived outline development.

Insert Figure 2 about here

Insert Table 1 about here

Method

Four test construction projects were selected for the evaluation of the computer-derived outline. Each team consisted of four subject-matter specialists (SMSs) from their respective career fields. The SMSs were either selectees or held the Air Force grade of E-7, Master Sergeant, or higher. These SMSs and the test psychologists who conducted each project voluntarily agreed to use the computer-derived outline procedure. The four projects were as follows: 631X0, Fuel Specialist and Fuel Supervisor; 316X0F, Missile Systems Analyst; 328X3, Electronic Warfare; and 701X0, Chapel Management. For each project a recently completed occupational survey was available. For the 328X3 and the 701X0, current computer programming also allowed for the presorting of the tasks by Specialty Training Standard (STS) and the prerating of the tasks according to the usability importance of the tasks for testing. An occupational survey task was selected for the computer-derived outline if twenty percent or more of the members performed the task. Supervisory tasks were not selected, nor were tasks selected with resulting testing importance values of zero (Phalen, 1978).

The first three test construction teams, 631X0, 316X0F, and 328X3 began with the conventional outline development procedure and then they evaluated and finalized the computer-derived outline. The fourth team, 701X0, developed only the computer-derived outline independently of the conventional outline procedure. Because of the relative consistency of conventional outlines from one test revision to the next, the previous team's conventional outline was used as the criterion for the 701X0 project. Verbal feedback was elicited from all of the SMSs. The 328X3 and 701X0 teams also completed the Outline Questionnaire (Fig 3).

Insert Figure 3 about here

Using the conventional outline as a criterion, the resultant percentage weights were compared for each major outline area. Percentage weight differences were computed between the conventional outline major areas and the major areas of the computer-derived outline. These differences were compared with the total number of tasks printed on the computer product. Homogeneity of tasks, i.e., the commonality of tasks across the career field, was considered.

Responses to the Outline Questionnaire were separated into positive, negative, or indifferent response to assess attitude toward the computer-derived outline. The last questionnaire item was used to assess SMS position as to which outline development procedure they preferred. Attitude toward the occupational survey (third questionnaire item) was compared with the preferred outline development procedure.

Results

For all four test construction projects the percentage weight differences for each major outline area between the final computer-derived outline and the conventional outline were not significantly different. The comparison for the 316X0F project is shown in Table 2. The resultant percentage weights for the computer-derived outline are presented for the initial computer product and also for the final outline. The number of tasks selected for each skill level is indicated in the footnote.

Insert Table 2 about here

The comparison for the 701X0 project is presented in Table 3. The large difference in major outline area III resulted from a low number of selected tasks that could not be referenced to the career development course (CDC).

Insert Table 3 about here

Table 4 is the 631X0 project comparison. The zero weight for the conventional and final computer-derived outline area V was a judgment decision by

Insert Table 4 about here

the team. They felt the tasks for area V were more appropriate in other areas.

The 328X3 project comparison is shown in Table 5. Even though the number of tasks selected for the 631X0 and the 328X3 projects was less than eleven percent of the total number of job inventory tasks, the teams were able to develop a final computer-derived outline. Because of the heterogeneity of the 328X3 career field due to the distinct differences in equipment from base to base, the team concluded that an additional major outline area (III) was needed. The new area was on basic principles which could be generalized across the career field.

Insert Table 5 about here

Table 6 shows the relationship of the total (absolute) percentage weight difference between projects. There was a general trend for the percentage weight differences to decrease as the total number of tasks selected increased (JIC=-.71, $p < .025$)[1]. In every case, the smallest percentage weight difference was between the conventional outline and the final computer-derived outline (JIC=.82, $p < .01$). This includes the 701X0 project during which the computer-derived outline was developed independently of the conventional outline.

Insert Table 6 about here

The attitude response from the Outline Questionnaire is presented in Table 7. There were nearly twice as many positive responses as negative (HLC=.57, $p < .005$)[2]. The response to the last item on the questionnaire indicated that six SMSs would choose the computer-derived outline procedure over the conventional outline. The remaining two SMSs were indifferent. A comparison of the attitude toward the occupational survey (third questionnaire item) with the preferred outline development procedure revealed a dichotomy. All four SMSs on the 701X0 team responded negatively to the third item and three of the four on the 328X3 team were not sure.

Insert Table 7 about here

Discussion

The purpose of this study was to evaluate the experimental application of the computer-derived test outline procedure. Conventional test outlines were used as a criterion reference during four actual test development projects. For one of these projects only a computer-derived outline was developed independently of the conventional outline procedure. For the other three projects, both types of outlines were developed. The resultant percentage weights for major outline areas were compared. In every case the smallest percentage weight difference was between the conventional outline and the final computer-derived outline. The differences were not significant and the relationship correlated .82 ($p < .01$). Overall, the more homogeneous the career field, the larger the numbers of occupational survey tasks selected for the initial computer product (correlation .71, $p < .025$). In conjunction with this trend, the more homogeneous the career field the smaller the percentage weight differences. This meant that the more homogeneous the career field, the closer the initial computer product was to reflecting the conventional outline. This relationship was also shown with the computer-derived outline developed independently of the conventional outline procedure. This finding substantiates an existing feeling in test construction. This feeling is that the problems involved in developing a test development plan decrease proportionly to the homogeneity of the career field. A compensation for more heterogeneous career fields is to decrease the task selection criterion from 20 percent to about 10 percent members performing. Even though the additional tasks will be performed by a small percentage of personnel, there are usually basic principles that can be generalized across the career field.

---

[1] Jenkins Index of Covariation (Jenkins & Hatcher, 1976)

[2] Hi-Lo Coefficient (Davidoff & Goheen, 1953)

On the Outline Questionnaire, there were nearly twice as many positive responses as negative ($p < .005$). The response to the last item on the questionnaire indicated that six out of the eight SMSs would choose the computer-derived outline development procedure over the conventional. The other two SMSs were indifferent. In comparison, the reaction to the occupational survey data indicated a dichotomy. Four SMSs responded negatively, three were indifferent, and one was positive. This comparison indicated that even though the SMSs indicated reluctance to fully accept the occupational survey as a true and complete picture of their career field, they recognized the advantage of using the occupational survey in the test development process. The computer-derived outline procedure caused the SMSs to become involved with the occupational survey data. The SMSs admitted that the survey data enhanced their ability to reach agreement on test content.

The four test construction teams felt that the computer product they used was easy enough to follow. They agreed that the Specialty Training Standard (STS) order was the logical format. Although an additional table was furnished the team to assist them in converting testing importance values into the suggested number of test questions for each task, this step was still too complex. The conversion needs to be incorporated into the computer program. Even then there will still be the need for the human element, i.e., the team's evaluation of the computer product.

Every team felt a need to readjust the tasks shown on the initial computer product and the resultant percentage weights for the major outline areas. This step was the most complicated with the more heterogeneous career fields. Yet, even with these adjustments, the relative time required to develop the computer-derived outline, ranging from one-half day to a day and a half, is no longer than the time used for development of the conventional outline. The incorporation into the computer program of the conversion from testing importance values to actual outline weights for each task will probably decrease the amount of time required for outline development.

Four factors play a key role in the feasibility of fully implementing the computer-derived outline procedure. The first is currency of the occupational survey data, i.e., does the occupational survey depict the current career field. The present surveying operation is closer to keeping up with career field changes than ever before. The second is timeliness of computer related support. Test development schedules are firm, so the necessary computer product must be available at the beginning of the project. Once the need is confirmed, it is possible to prepare the computer product well in advance of a project. The third is the workload on personnel, i.e., being able to do the job within existing resources. Existing test support activities should be evaluated to determine how the present support procedures could be altered to fit the new outline development procedures without increasing the workload on personnel. The fourth factor is the SMS attitude toward occupational survey data. The briefing to the SMSs about the survey should include a discussion of quality control measures taken by the occupational survey activity to insure valid data. Also, steps should be taken to insure that each person completing a job inventory for occupational survey understands the importance of accurate responses.

Since each test development team felt the need to readjust task distribution and percentage weights on the initial computer product for their final outline, there is a need for further refinement of the computer-derived outline procedure. The evaluation should include further validation of the formula used to compute the testing importance values.

As a result of this study it can be concluded that the computer-derived outline procedure is viable for test construction. The computer product is in a format that is agreeable to the SMSs that have used it. The procedure for using the computer product can be followed even by the individual who is not acquainted with occupational survey data. The final computer-derived outline is not significantly different from the time-tested conventional outline. However, the computer-derived outline does directly incorporate occupational survey data into test development procedures. The incorporation of occupational survey data expands the input for test outline development from four SMSs to the field of survey respondents. This expansion strengthens the content validity position of the resultant test. The strength of the computer-derived outline along with the feasibility of the procedure shown in this study indicate that the Occupational Measurement Center should proceed to incrementally implement the computer-derived outline procedure with concurrent evaluation.

References

Davidoff, M. D. and Goheen, H. W., *Psychometrica*, 1953, *18*, 115-121.

Equal Employment Opportunity Coordinating Council. *Uniform Guidelines on Employment Selection Procedures* (draft). October 18, 1977.

Jenkins, W. O. and Hatcher, N. C. *The Design of Behavioral Experiments*. Auburn University at Montgomery, AL (unpublished manuscript), 1976, 323.

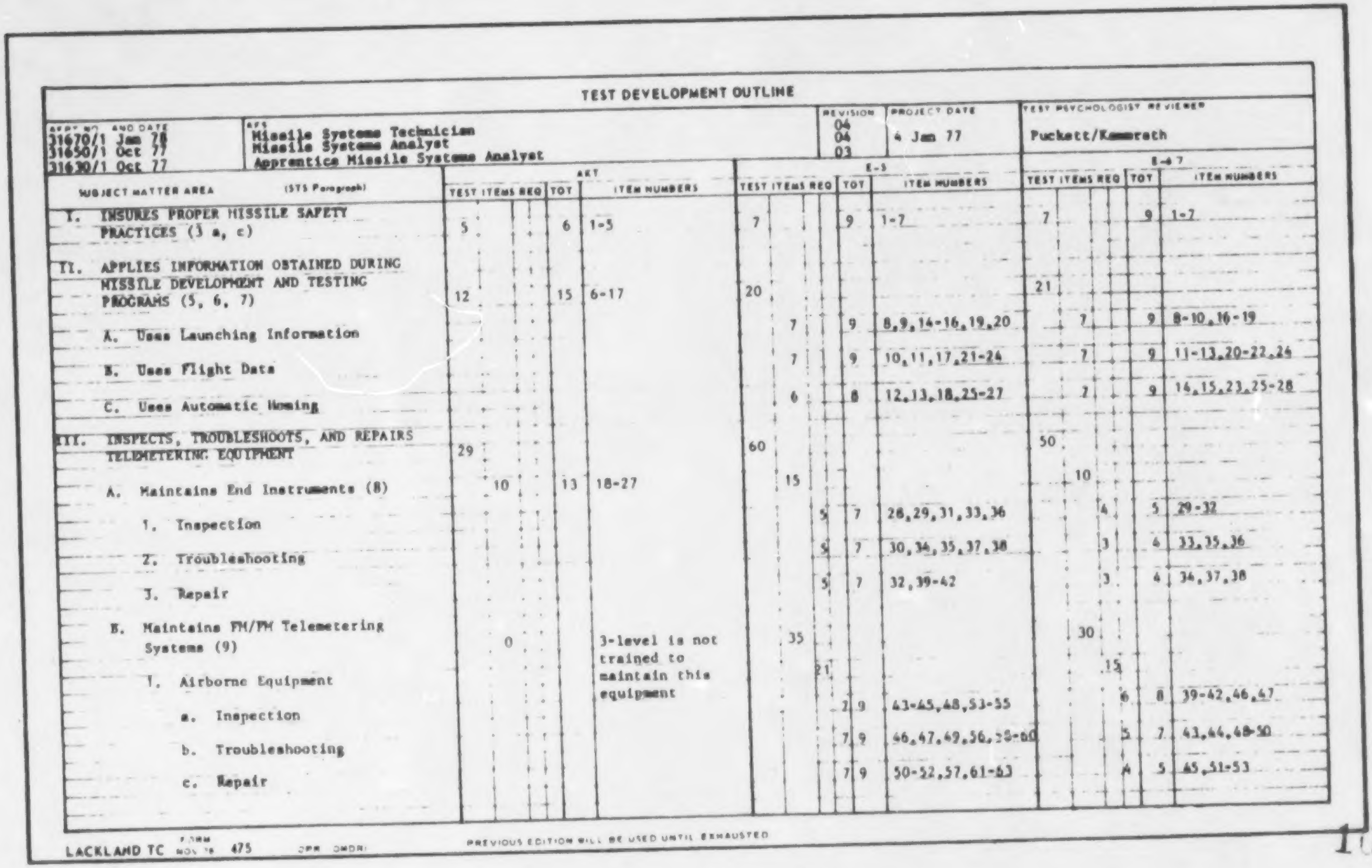
Phalen, W. J. The Development of a Technique for Using Occupational Survey Data to Construct and Weight Computer-Derived Test Outlines for Air Force Specialty Knowledge Tests (SKTs). *Proceedings of the 20th Annual Conference of the Military Testing Association*, 1978.

USAF Occupational Measurement Center. *Handbook for Construction of the SKT and Associated Tests*. Lackland AFB, TX: author, June 1, 1977.

USAF Occupational Measurement Center. *Subject-Matter Specialist Handbook for Construction of the SKT and Associated Tests*. Lackland AFB, TX: author, July 1, 1977.

Vaughan, D. S. Prediction of Test Outline Weights from Occupational Survey Data. *Proceedings of the 18th Annual Conference of the Military Testing Association*, 1976, 435-460.

Vaughan, D. S. The Interface Between Occupational Survey and Test Construction. *Proceedings of the 19th Annual Conference of the Military Testing Association*, 1977, 798-800.

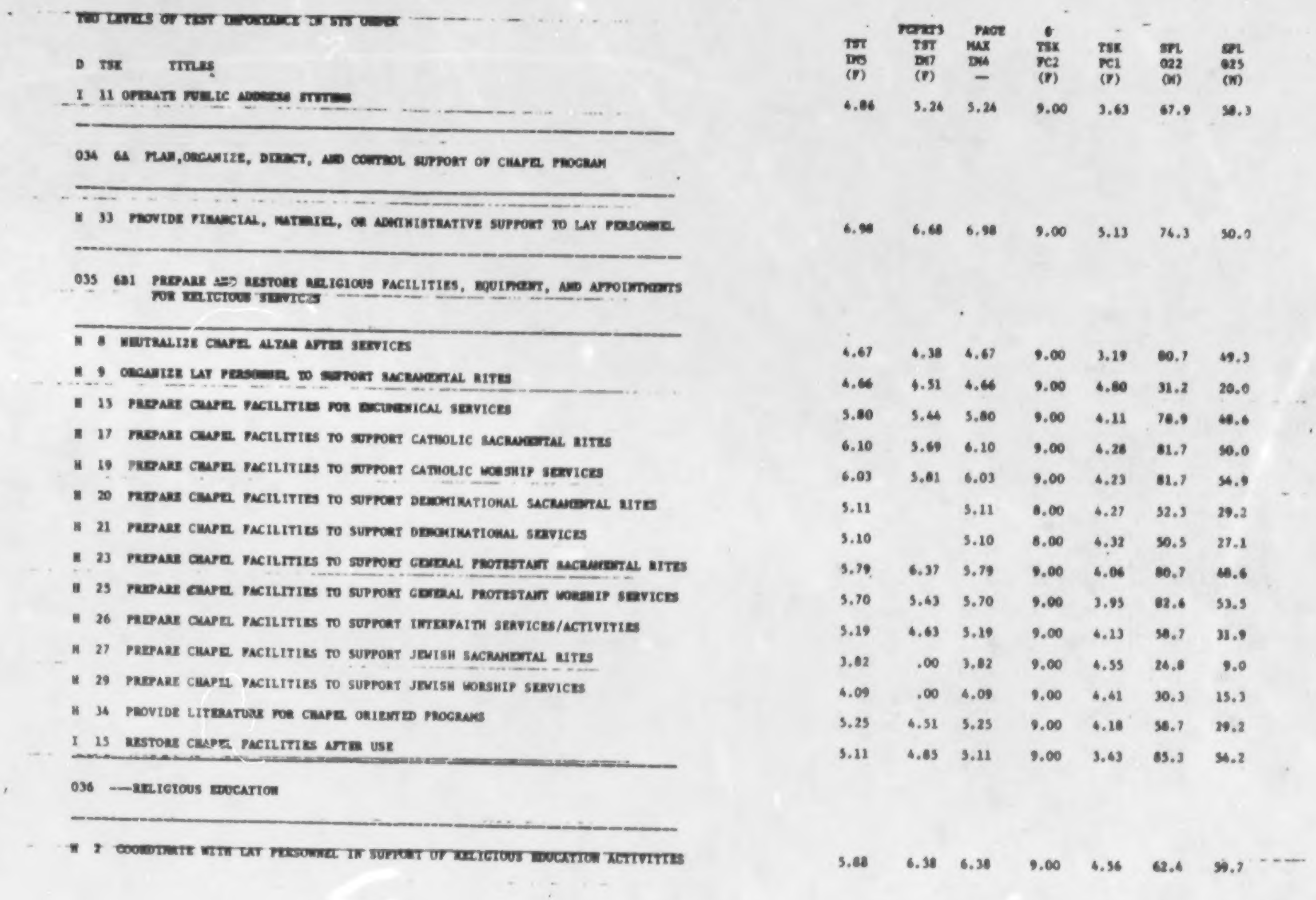


TEST DEVELOPMENT OUTLINE

AFPT NO. AND DATE 31670/1 Jan 78 31650/1 Oct 77 31630/1 Oct 77	AFS Missile Systems Technician Missile Systems Analyst Apprentice Missile Systems Analyst			REVISION 04 04 03	PROJECT DATE 4 Jan 77		TEST PSYCHOLOGIST REVIEWER Puckett/Kammrath		
	AKT			E-5			E-6 7		
SUBJECT MATTER AREA (STS Paragraph)	TEST ITEMS REQ	TOT	ITEM NUMBERS	TEST ITEMS REQ	TOT	ITEM NUMBERS	TEST ITEMS REQ	TOT	ITEM NUMBERS
I. INSURES PROPER MISSILE SAFETY PRACTICES (3 a, c)	5	6	1-5	7	9	1-7	7	9	1-7
II. APPLIES INFORMATION OBTAINED DURING MISSILE DEVELOPMENT AND TESTING PROGRAMS (5, 6, 7)	12	15	6-17	20			21		
A. Uses Launching Information				7	9	8,9,14-16,19,20	7	9	8-10,16-19
B. Uses Flight Data				7	9	10,11,17,21-24	7	9	11-13,20-22,24
C. Uses Automatic Homing				6	8	12,13,18,25-27	7	9	14,15,23,25-28
III. INSPECTS, TROUBLESHOOTS, AND REPAIRS TELEMETERING EQUIPMENT	29			60			50		
A. Maintains End Instruments (8)	10	13	18-27	15			10		
1. Inspection				5	7	28,29,31,33,36	4	5	29-32
2. Troubleshooting				5	7	30,34,35,37,38	3	4	33,35,36
3. Repair				5	7	32,39-42	3	4	34,37,38
B. Maintains FM/PM Telemetering Systems (9)	0		3-level is not trained to maintain this equipment	35			30		
1. Airborne Equipment				21			15		
a. Inspection				7	9	43-45,48,53-55	6	8	39-42,46,47
b. Troubleshooting				7	9	46,47,49,56,58-60	5	7	43,44,48-50
c. Repair				7	9	50-52,57,61-63	4	5	45,51-53

LACKLAND TC FORM NOV 76 475 OPR DMDR PREVIOUS EDITION WILL BE USED UNTIL EXHAUSTED

Figure 1. Conventional Outline



TWO LEVELS OF TEST IMPORTANCE IN STS ORDER

D	TSK	TITLES	TST IM5 (F)	EXPRTS TST IM7 (F)	PAGE MAX IM4 —	6 TSK PC2 (F)	TSK PC1 (F)	SPL 022 (M)	SPL 025 (M)
I	11	OPERATE PUBLIC ADDRESS SYSTEMS	4.86	5.24	5.24	9.00	3.63	67.9	58.3
034	6A	PLAN,ORGANIZE, DIRECT, AND CONTROL SUPPORT OF CHAPEL PROGRAM							
H	33	PROVIDE FINANCIAL, MATERIEL, OR ADMINISTRATIVE SUPPORT TO LAY PERSONNEL	6.98	6.68	6.98	9.00	5.13	74.3	50.0
035	6B1	PREPARE AND RESTORE RELIGIOUS FACILITIES, EQUIPMENT, AND APPOINTMENTS FOR RELIGIOUS SERVICES							
H	8	NEUTRALIZE CHAPEL ALTAR AFTER SERVICES	4.67	4.38	4.67	9.00	3.19	80.7	49.3
H	9	ORGANIZE LAY PERSONNEL TO SUPPORT SACRAMENTAL RITES	4.66	4.51	4.66	9.00	4.80	31.2	20.0
H	15	PREPARE CHAPEL FACILITIES FOR ECUMENICAL SERVICES	5.80	5.44	5.80	9.00	4.11	78.9	48.6
H	17	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC SACRAMENTAL RITES	6.10	5.69	6.10	9.00	4.28	81.7	50.0
H	19	PREPARE CHAPEL FACILITIES TO SUPPORT CATHOLIC WORSHIP SERVICES	6.03	5.81	6.03	9.00	4.23	81.7	54.9
H	20	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SACRAMENTAL RITES	5.11		5.11	8.00	4.27	52.3	29.2
H	21	PREPARE CHAPEL FACILITIES TO SUPPORT DENOMINATIONAL SERVICES	5.10		5.10	8.00	4.32	50.5	27.1
H	23	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT SACRAMENTAL RITES	5.79	6.37	5.79	9.00	4.06	80.7	48.6
H	25	PREPARE CHAPEL FACILITIES TO SUPPORT GENERAL PROTESTANT WORSHIP SERVICES	5.70	5.43	5.70	9.00	3.95	82.6	53.5
H	26	PREPARE CHAPEL FACILITIES TO SUPPORT INTERFAITH SERVICES/ACTIVITIES	5.19	4.63	5.19	9.00	4.13	58.7	31.9
H	27	PREPARE CHAPEL FACILITIES TO SUPPORT JEWISH SACRAMENTAL RITES	3.82	.00	3.82	9.00	4.55	24.8	9.0
H	29	PREPARE CHAPEL FACILITIES TO SUPPORT JEWISH WORSHIP SERVICES	4.09	.00	4.09	9.00	4.41	30.3	15.3
H	34	PROVIDE LITERATURE FOR CHAPEL ORIENTED PROGRAMS	5.25	4.51	5.25	9.00	4.18	58.7	29.2
I	15	RESTORE CHAPEL FACILITIES AFTER USE	5.11	4.85	5.11	9.00	3.63	85.3	56.2
036		---RELIGIOUS EDUCATION							
H	2	COORDINATE WITH LAY PERSONNEL IN SUPPORT OF RELIGIOUS EDUCATION ACTIVITIES	5.88	6.38	6.38	9.00	4.56	62.4	59.7

Figure 2. Sample Page of Initial Computer Product

AFSC________________ TEST PROJECT DATE________________________

OUTLINE QUESTIONNAIRE

Indicate the degree to which you agree with the following statements:

Agree Disagree

______I feel the computer-derived outline was easier to develop than the conventional outline. A B C D E

______In comparison with the conventional outline I feel the computer-derived outline more accurately reflects the true job situation in the field. A B C D E

______I have confidence that the survey data used to compile the computer outline is accurate and dependable. A B C D E

______I feel the format of the computer outline is difficult to understand. A B C D E

______I found the computer outline product very easy to work with. A B C D E

______I feel that SKT-usable references are available for all areas listed in the computer outline. A B C D E

______I found that the computer product, as printed, sufficiently covered all STS areas. A B C D E

______I found that a substantial amount of information had to be added before the computer product could be used as a test outline. A B C D E

______Given a choice, I would prefer to develop an outline from the computer product rather than use the conventional method. A B C D E

Figure 3. Outline Questionnaire

Table 1

Conversion Table for Determining Equivalent Percentage Weights from 328X3 Testing Importance Value

Testing Importance Value	Percentage Weight			
	5-Skill Level		7-Skill Level	
7.5	5.3	5	8.3	8
7.0	4.9		7.7	
6.5	4.6		7.2	7
6.0	4.2	4	6.6	
5.5	3.8		6.1	6
5.0	3.5		5.5	
4.5	3.2	3	5.0	5
4.0	2.8		4.4	4
3.5	2.5		3.9	
3.0	2.1	2	3.3	3
2.5	1.8		2.8	
2.0	1.4	1	2.2	2
1.5	1.1		1.7	
Total Testing Imp Value	142.59		90.53	

1040

Table 2

316X0F Percentage Weight Comparison of Conventional Outline with Computer-Derived Outline*

Skill Level	Major Outline Area	Conventional Outline % (A)	Computer-Derived Outline % (B=Initial)	Computer-Derived Outline % (C=Final)	Percentage Weight Differences (B-A)	Percentage Weight Differences (C-A)	Percentage Weight Differences (C-B)
5	I	46	32	48	-14	2	16
	II	20	21	9	1	-11	-12
	III	34	47	43	13	9	- 4
Total (Absolute)		100	100	100	28	22	32
7	I	46	32	49	-14	3	17
	II	20	22	9	2	-11	-13
	III	34	46	42	12	8	- 4
Total (Absolute)		100	100	100	28	22	34

*Total number job inventory tasks: 783
Total number tasks selected 5-level: 143
Total number tasks selected 7-level: 140

Table 3

701X0 Percentage Weight Comparison of Conventional Outline with Computer-Derived Outline*

Skill Level	Major Outline Area	Conventional Outline %	Computer-Derived Outline %		Percentage Weight Differences		
		(A)	(B=Initial)	(C=Final)	(B-A)	(C-A)	(C-B)
5	I	12	18	25	6	13	7
	II	3	2	2	- 1	- 1	0
	III**	9	37	10	28	1	-27
	IV	34	19	30	-15	- 4	11
	V	40	24	33	-16	- 7	9
	VI	2	0	0	- 2	- 2	0
Total (Absolute)		100	100	100	68	28	54
7	I	26	24	30	- 2	4	6
	II	1	1	0	0	- 1	- 1
	III**	4	44	5	40	1	-39
	IV	40	23	35	-17	- 5	12
	V	27	8	30	-19	3	22
	VI	2	0	0	- 2	- 2	0
Total (Absolute)		100	100	100	80	16	80

*Total number job inventory tasks: 216
Total number tasks selected 5-level: 62
Total number tasks selected 7-level: 48

**Although high percent members performing, low number of tasks could be referenced to the Career Development Course (CDC)

Table 4

631X0 Percentage Weight Comparison of Conventional Outline with Computer-Derived Outline*

Skill Level	Major Outline Area	Conventional Outline % (A)	Computer-Derived Outline % (B=Initial)	(C=Final)	Percentage Weight Differences (B-A)	(C-A)	(C-B)
5	I	52	29	46	-23	- 6	17
	II	25	21	31	- 4	6	10
	III	13	13	17	0	4	4
	IV	10	29	6	19	- 4	-23
	V**	0	8	0	8	0	- 8
Total (Absolute)		100	100	100	54	20	62
7	I	51	32	44	-19	- 7	12
	II	26	19	33	- 7	7	14
	III	15	15	20	0	5	5
	IV	8	18	3	10	- 5	-16
	V**	0	16	0	16	0	-16
Total (Absolute)		100	100	100	52	24	62

*Total number job inventory tasks: 374
Total number tasks selected 5-level: 38
Total number tasks selected 7-level: 26
**Team felt tasks were more appropriate under a separate heading

Table 5

328X3 Percentage Weight Comparison of Conventional Outline with Computer-Derived Outline*

Skill Level	Major Outline Area	Conventional Outline % (A)	Computer-Derived Outline % (B=Initial)	Computer-Derived Outline % (C=Final)	Percentage Weight Differences (B-A)	Percentage Weight Differences (C-A)	Percentage Weight Differences (C-B)
5	I	4	21	14	17	10	- 7
	II	32	79	37	47	5	-42
	III**	64	0	49	-64	-15	49
Total (Absolute)		100	100	100	128	30	98
7	I	11	24	24	13	13	0
	II	39	76	48	37	9	-28
	III***	50	0	28	-50	-22	28
Total (Absolute)		100	100	100	100	44	56

*Total number job inventory tasks: 768
Total number tasks selected 5-level: 27
Total number tasks selected 7-level: 17
**Team desired a basic principles area

Table 6

Comparison By Project of Total (Absolute) Percentage Weight Differences Between Conventional Outline (A) and Computer-Derived Outline (B=Initial; C=Final)*

	5-Skill Level Percentage Weight Difference			Total # Tasks Selected	7-Skill Level Percentage Weight Difference			Total # Tasks Selected
	(B-A)	(C-A)	(C-B)		(B-A)	(C-A)	(C-B)	
316X0F	28	22	32	142	28	22	34	140
701X0	68	28	54	62	80	16	80	48
631X0	54	20	62	38	52	24	62	26
328X3	128	30	98	27	100	44	56	17
Range of Percentage Difference	100	10	66		72	28	46	

*C-A differences not significant. Generally as total tasks selected increases, percentage weight differences decrease (correlation -.71, p<.01). Smallest percentage weight difference C-A (correlation .82, p<.01).

Table 7

Positive (+), Indifferent (0), and Negative (-) Attitudes Toward the Computer-Derived Outline Procedures Based on Responses to the Outline Questionnaire*

Attitude Response	Test Development Project 328X3	701X0	Sum
+	21	10	31
0	10	14	24
-	5	12	17
Sum	36	36	72

*Significant difference between positive and negative (p<.005).

A Generalization of Sequential Analysis to Decision Making with Tailored Testing

by

Mark D. Reckase
University of Missouri-Columbia

During the last decade, there has been increasing interest in the individualization of instruction and the maintenance of high standards of quality in the students graduated from instructional programs. Both individualization and the maintenance of quality require achievement measurement procedures that can accurately determine whether a student is above or below a pre-set criterion score. Also, the relatively new areas of criterion-referenced measurement and mastery learning programs require accurate procedures for classification into the two groups (pass and fail) for their operation. If the classification can be done quickly with only a few test items, this would be a desirable attribute for a procedure.

Most decision procedures described in the current literature are based on sampling a fixed number of test items for a domain and using either classical or Bayesian decision rules for determining a person's position relative to a criterion (see Millman, 1974 and Hambleton, Swaminathan, Algina and Coalson, 1978 for reviews of these techniques). However, a family of procedures exists that has been shown to yield a smaller expected sample size for testing many hypotheses while holding the power of the test at the same level as the fixed sample size procedures (Wald, 1947). These are sequential procedures that have the characteristics of taking single observations and deciding after each observation if a classification should be made or if more information is needed--that is, if another observation should be taken. For many classifications, sequential procedures have been proven to be much more efficient than fixed sample size procedures by exhibiting high accuracy with relatively small sample sizes (Wald, 1947).

A simple example will be used to show how the number of test items used for classifying a student can be reduced while the accuracy of classification stays the same as for a full length test. Suppose a student who has not mastered the material from a unit of instruction is given a ten item quiz for the purpose of diagnosing that fact. Suppose further that an 80% criterion has been set for success. The usual procedure would be to give the ten item quiz, score it, and if the score were seven or less, give remedial instruction. When using

---

Paper presented at the meeting of the Military Testing Association, Oklahoma City, Oklahoma, October 30-November 2, 1978. This research was supported by contract number N00014-77-C-0097 from the Personnel and Training Research Programs of the Office of Naval Research.

1047

a sequential procedure, items would be administered one at a time and testing would stop as soon as three items were missed. The largest number of items administered would be ten, so the average number administered must be less then ten, but the same classification criterion has been used for both testing procedures.

A particular sequential procedure that has been applied to measurement decision problems in the past and that has shown promise for the future is the sequential probability ratio test (SPRT) developed by Wald (1947). This procedure has been thoroughly analysed within the mathematical statistics framework (Govindarajulu, 1975) and has recently been rediscovered by measurement theorists (Sixtl, 1974; Epstein, 1978). In this paper the SPRT will be generalized to tailored testing applications. However, a brief description of this sequential decision model will be given first.

The Sequential Probability Ratio Test (SPRT)

The sequential probability ratio test was originally developed to determine which of two population parameter values is most likely true for a given set of data. For example, one might be interested in determining whether the proportion failing a criterion-referenced test is more likely .5 or .8. If a certain three of five students sampled from a population fail to exceed a criterion, this event would have a probability of $.5^5 = .03125$ if the .5 hypothesis were correct and $.8 \times .8 \times .8 \times .2 \times .2 = .02048$ if the .8 hypothesis were correct. The question now becomes whether the difference in these two probabilities is great enough to select the .5 hypothesis over the .8 hypothesis.

To make this decision, Wald took the ratio of the two probabilities, $\frac{.02048}{.03125} = .69526$. If the ratio were sufficiently larger than 1.0, the .8 probability would be accepted as correct. If it were much smaller than 1.0, the 15 probability would be considered as correct. Note that for the sequential procedure, this ratio would be computed after each observation, and a decision concerning the .5 or .8 parameter would be made as soon as the ratio passed either an upper or lower cutoff value.

To totally specify the SPRT procedure some means must be given to determine the two cutoff values, Θ_0 and Θ_1. These cutoffs are directly dependent on the error rates that are deemed acceptable in choosing between the two parameter values. The probability of choosing .8 when .5 is true is defined as an α error and the probability of choosing .5 when .8 is true is defined as a β error. Wald has shown that these error rates will be at least as low as the values of α and β when the two cutoff values are set at:

$$\text{lower cutoff} = B = \frac{(1 - \beta)}{\alpha}$$

$$\text{upper cutoff} = A = \frac{\beta}{(1 - \alpha)}$$

If α is set at .02 and β at .1 the cutoff values are A = 45 and B = .102. If after each observation the ratio of the probabilities is more extreme than either test value, the appropriate parameter value is accepted as true. If it is between the two cutoffs another observation is taken.

Testing the hypothesis that .5 is correct against the hypothesis that .8 is correct is seldom of interest in a criterion referenced testing setting. A more common hypothesis is that a person is below a cutoff value as opposed to being above the value. Wald has shown that this complex hypothesis can be tested in the same way as the two simple hypotheses by selecting a cutoff value and then specifying a region of indifference around the cutoff in which the classification as to above or below the cutoff is equally good. The lower end of the indifference region is used as the lower simple hypothesis, H_0, while the upper end of the region is used as the upper simple hypothesis, H_1. The A and B values used in the significance test are determined in the same manner as above.

An example of testing this type of complex hypothesis in a criterion-referenced testing situation can be given as follows. Suppose we want to determine if a student can answer 90% of the items in an item domain. Suppose also that we are indifferent as to whether they are classified as high or low in the region from 89% to 95%. We would then randomly select items one at a time from the domain and determine the probability of the response strings under the H_0: $\pi = .89$ and H_1: $\pi = .95$. The ratio of the probabilities of the response strings would be computed as previously described and then compared to the A and B decision values. If the ratio were below the B value, the person would be classified as below the criterion; if it were above the A value, the person would be classified as above the criterion. If the ratio were between A and B, another item would be administered.

Note that in this example items were randomly sampled one at a time without replacement and then administered. This is called a sequential random sample and it is one of the basic assumptions used in deriving the method.

Description of the Characteristics of SPRT

When using a SPRT for decision making, two functions are derived to describe the accuracy and efficiency of the procedure. The first is called the operating characteristic (OC) function of the test.

1049

The OC function gives the probability of accepting the null hypothesis as a function of the unknown parameter of interest, $\hat{\theta}$. For criterion referenced testing, the null hypothesis is usually that the examinee is below the criterion. Typically the plot of this function is an S-shaped curve asymptoting at 1.0 one the left and 0.0 on the right (see Figure 1). Wald has shown that at the lower simple hypothesis value, Θ_0, the curve will have a height of $1 - \alpha$, while at the upper critical value, Θ_1, the height is β. The slope of the function between these two points is dependent on the width of the indifference region--the wider it is, the flatter the slope. Finally, the point of inflection of the curve is usually near the decision point. An ideal OC curve would approximate a step function, dropping abruptly from a probability of 1.0 of accepting the null hypothesis below the decision criterion to a probability of 0.0 of accepting the null hypothesis immediately above the criterion.

Insert Figure 1 about here

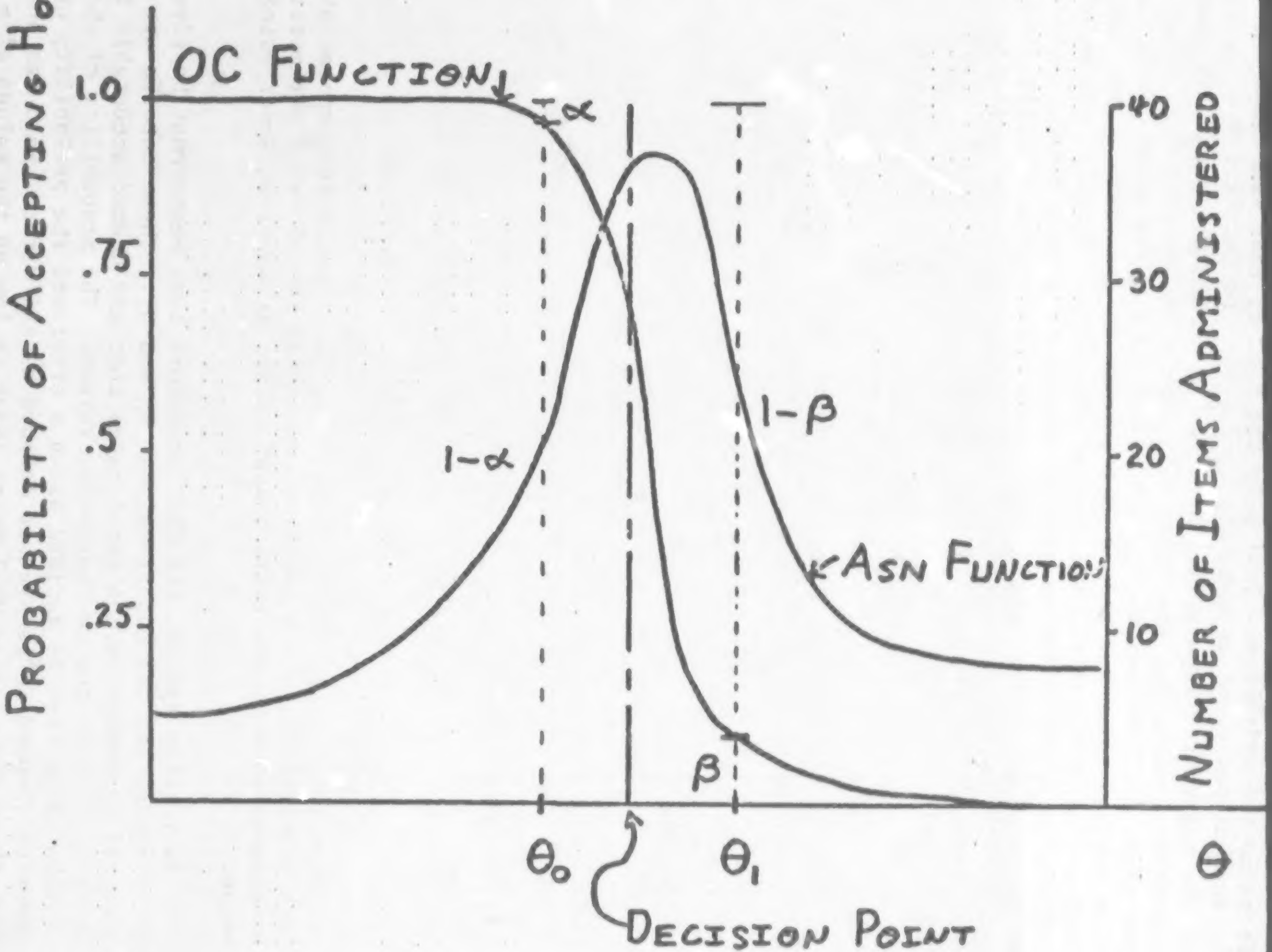
The second function used to evaluate the operation of the SPRT decision rule is the average sample number (ASN) function. This ASN function gives the average number of observations required to make a decision as a function of the variable used to make the decision. This function typically plots as a unimodal curve with its mode near the decision point (see Figure 1). The curve asymptotes to zero in either direction from the mode. Since this function gives an indication of how many observations are required for a decision, the lower the modal value the better. That is, corresponding to a given OC function, we would like the ASN function to be as low as possible throughout the variable range, indicating that only a few observations are required. Another desirable feature for an ASN function is a quick decline from the mode, indicating that decisions require few observations if a person is not near the decision point.

The magnitudes of the values of these two functions are related to each other. As the slope of the OC function increases, the values of the ASN function will usually increase. If a flatter OC is acceptable, the values of the ASN function will be less. In using a SPRT, a compromise must be reached between precision (as shown by the OC curve) and sample size (as shown by the ASN function). Both of these functions will be used to evaluate the SPRT for use with tailored testing.

Generalization of the SPRT to Tailored Testing

As mentioned above, the SPRT as developed by Wald assumes that observations are taken using a sequential random sample. In a criterion-referenced test, this would mean that items would be selected and administered at random one at a time from a domain of items. Although random sampling is philosophically acceptable with criterion-referenced

FIGURE 1
OC AND ASN FUNCTIONS



testing, it is at odds with the purposes of tailored testing. In this latter case, the purpose is to select items to match the abilities of each pupil rather than to administer a random selection. As a result of matching items to pupils, the testing situation should be more efficient and accurate. Since the purpose of this paper is to merge the SPRT procedure with tailored testing, an initial task is to determine whether the sequential random sample assumption is really necessary.

A detailed analysis of Wald's (1947) work by the present author indicates that the assumption was only needed to make the derivation of the OC and ASN functions possible. Without the assumption, the characteristics of each item must be specified, resulting in many nuisance parameters that cannot be eliminated. However, the test statistics still operate in the same way, so the procedure can still be used. The OC and ASN functions will be developed using simulations in this paper since they cannot be developed using the usual formulas. An example will now be given showing the use of the SPRT with tailored testing.

Suppose it is desirable to determine whether a student's performance on a mod[illegible]e of instruction is above or below a pre-set criterion score. Since [illegible]rigin of the latent trait ability scale is arbitrary, the criterion score can be set at 0.0 without loss of generality. An indifference region must now be specified for this criterion score. Assume that ability estimates in the region around 0.0 have been found to have a standard error of .3 for the population and item pool of interest. Therefore, the indifference region will be specified as -.3 to +.3, and $\Theta_0 = -.3$ and $\Theta_1 = +.3$ are used for the SPRT.

Next, the acceptable error rates for the classification decision must be specified. For this decision suppose it was felt to be a more serious error to classify a person above the criterion score when they should have been classified low, than to classify below when they should have been above. Therefore α was set at .02 and β at .1 and two classification values for the SPRT would then be A = 45 and B = .102.

With the specification of this preliminary information, the operation of the SPRT can begin. When no previous information is available about a student, the tailored testing procedure first administers an an item of moderate difficulty. Using a one parameter logistic model, this first item has a difficulty value, b, of 0.0. Suppose the student gives a correct response to this item. The probability of this response under $\Theta_0 = -.3$ is given by

$$P_{01} = P_1(\Theta_o) = \frac{e^{(\Theta_o - b_1)}}{1 + e^{(\Theta_o - b_1)}} = \frac{e^{(-.3 - 0)}}{1 + e^{(-.3 - 0)}} = .426$$

where p_{01} is the probability of the response after one item under H_o, and the formula is that of the one-parameter logistic model. The probability under $\Theta_1 = .3$ is given by

$$P_{11} = P_1(\Theta_1) = \frac{e^{(\Theta_1 - b_1)}}{1 + e^{(\Theta_1 - b_1)}} = \frac{e^{(.3 - 0)}}{1 + e^{(.3 - 0)}} = .574$$

where p_{11} is the probability of the response after one item under H_1. The value of the SPRT is given by

$$\frac{P_{11}}{P_{01}} = \frac{.574}{.426} = 1.347$$

Since this value is between A and B, no decision can be made and another item should administered. Since the first item was responded to correctly, a more difficult item will now be administered to try to match the person's ability, say an item of +.7 difficulty. If an incorrect response is obtained to this second item, the probability of the 1, 0 response string under Θ_o is

$$P_{02} = \prod_{i=1}^{2} P_i(\Theta_o)^{X_i} Q_i(\Theta_o)^{1 - X_i}$$

$$P_{02} = .426 \times .731 = .341$$

where p_{02} is the probability of the response string after two items, given Θ_o; $P_i(\Theta_o)$ is the probability of a correct response to Item i, given Θ_o; $Q_i(\Theta_o)$ is the probability of an incorrect response, and X_i is the response to Item i, (0 or 1).

Under Θ_1, the probability of the response string is

$$P_{12} = \prod_{i=1}^{2} P_i(\Theta_1)^{X_i} Q_i(\Theta_1)^{1 - X_i}$$

$$= .574 \times .401 = .230$$

The SPRT is then equal to

$$\frac{P_{12}}{P_{02}} = \frac{.230}{.341} = .674$$

Since this value is still between A and B, no decision can be made and a third item should be administered. The procedure would then continue in the same way until the ratio is more extreme than A or B. At the point, the appropriate decision would be made and testing would stop. In theory a very large number of items could be administered before a decision is made--although Wald has proven that the number is finite. However, in practice some reasonable upper limit is set on the number of items administered, 20 for example, and a decision is made after the twenty items on the basis of whether the probability ratio is above or below 1.0. This procedure is called a truncated SPRT.

As mentioned earlier, one of the assumptions of the SPRT is a sequential random sample. Since that assumption is not met, and also since in real situations, the procedure may be truncated, it is impossible to derive the ASN and OC functions. Therefore, the major purpose of this paper was to determine these functions through simulations and use this information to evaluate the procedures for use with criterion-referenced tailored testing.

Method

The OC and ASN functions were determined for tailored tests using both the one- and three-parameter logistic models based on maximum likelihood estimation of abilities. The three-parameter logistic model is an extension of the one-parameter model that includes discrimination and guessing parameters (See Lord and Novick, 1968, for further information). Simulations were used in both cases. The tailored testing procedures used have been described in detail by Koch and Reckase (1978), so they will not be described again here. However, to distinguish the techniques from other procedures, it should be stated that the procedures begin with an item of average difficulty and operate on a fixed step-size procedure until a correct and incorrect response is present. At that point a maximum likelihood ability estimate is obtained and the next item is selected to yield maximum information for that ability estimate. The procedures terminate when appropriate items are no longer available or if twenty items have been administered, whichever occurs first.

The simulated tailored testing procedures were identical to those described above, except that a random number generator replaced the human examinee. At the beginning of each simulation run the true ability of the simulated examinee was input into the program. This value was used to determine the true probability of a correct response to the administered items based on the model used, (one- or three-parameter logistic) and the estimated item parameters. A number was then randomly selected from a uniform distribution on the range from 0 to 1. If the selected number was less than the probability of a correct response, a correct response was recorded; otherwise an incorrect response was assigned. This procedure continued for each item in the tailored test.

Tailored tests were administered twenty-five times at each true ability using different seed numbers for the random number generator.

True abilities from -3 to +3 at .25 intervals were used for the one- and three-parameter models to evaluate the SPRT. Indifference regions of ±.3, ±.8, and ±1 were used in the evaluation. All simulations used the item parameters from a pool of 72 vocabulary items. This item pool had an approximately normal distribution of difficulty parameters.

During the administration of the tailored tests, probability ratios were computed after each item was administered. A decision was made to classify a person above or below the cutoff by comparing the SPRT value to an A value of 45 and a B value of .102, determined using $\alpha = .02$ and $\beta = .10$. A classification was made the first time these limits were exceeded. If the limits were not exceeded before the termination of the test, values above 1.0 were classified as above and the values below 1.0 were classified as low. At each true ability used for the simulation, the proportion of the 25 administrations classified low and the average number of items administered were computed. Plots of these values against the true abilities approximate the OC and ASN functions, respectively.

Results

The results of this research will be described in two parts; one for the one-parameter logistic model, and the other for the three-parameter logistic model. The plots of the OC and ASN functions summarize the results of the SPRT for these models.

Figure 2 shows the OC functions for the one-parameter logistic model based on the vocabulary item pool. The figure shows three graphs, one for each of the ±.3, ±.8, and ±1 indifference regions. Note that the curves are reasonably similar regardless of the indifference region. The similarity indicates that in all four cases the classification accuracy is nearly the same.

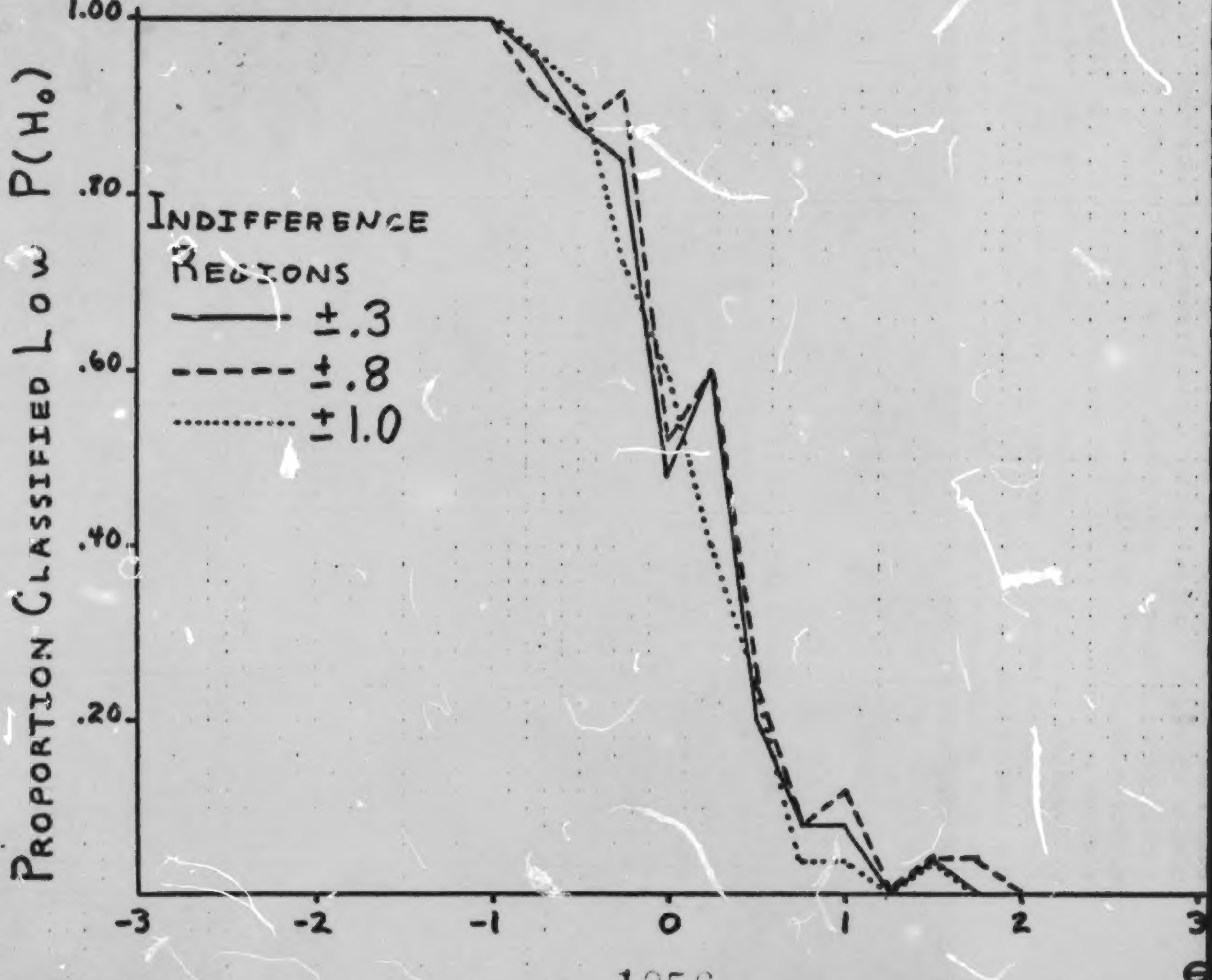
Insert Figure 2 about here

The values of the curves at the limits of the indifference region give further evaluative information. At the lower point, the OC function should pass through $1 - \alpha$. At the -.3 value, the curve is in fact .85 when it should be .98 showing the degrading effects of restrictive stopping rules used by the tailored testing procedure. At the -.8 and -1 points for the corresponding curves, the results are about as expected, being .94 and 1.00 rather than .98.

At the upper limit of the indifference region the OC function should have a value of .1. For the .3 case it is in fact .5 rather the .1, again showing the effects of truncating the procedure. At the values of .8 and 1, the values of the OC function were near or better than what they should have been based on the theoretically expected results.

FIGURE 2

ONE-PARAMETER OC FUNCTIONS FOR THREE INDIFFERENCE REGIONS



The ASN functions for the one-parameter model are given in Figure 3. The curves plotted correspond to the ASN functions using indifference regions for +.3, +.8, and +1. It can immediately be seen from the graph that there is a substantial difference in the average number of items needed to reach a decision, with the greatest number required when the indifference region is narrowest. It can also be seen that the largest expected number of items is near the decision point of 0.0 and that the average number drops off at the extreme abilities. The slight lack of symmetry in the curves is due to the fact that α was not equal to β. For abilities beyond +1, an average of only about 3 to 5 items was needed for classification for the wider regions, while 6 to 11 were needed for the +.3 indifference region. Note that the +.3 curve is approaching the arbitrary twenty item limit for the tailored tests, possibly reducing its magnitude.

Insert Figure 3 about here

Figure 4 shows the theoretical curves for the ASN and OC functions based on the +.3 indifference region for comparison purposes. An infinite number of items with difficulty 0.0 was assumed for the theoretical functions and the tests were assumed to have no upper limit on the number of items administered. A comparison of Figures 2 and 3 with Figure 4 shows that the OC curve for the theoretical function is steeper at the cutting point than the simulated curves and the ASN function is substantially higher. The difference in the theoretical and simulated OC curves shows the effect of the tailored testing stopping rule.

Insert Figure 4 about here

The results of the simulation of the three-parameter logistic tailored test are given in Figures 5 and 6. Figure 5 presents the OC functions for the three-parameter model, again using the indifference regions of +.3, +.8, and +1. Notice that, as with the one-parameter model, the OC curves are fairly similar for the three indifference regions throughout most of the range of ability. However, there are discrepancies for the +1.0 indifference range curve near the +1 and -1 points indicating a decline in decision precision for that region. At the -.3 value for the +.3 indifference range, the value of the curve is .96, fairly close to the .98 theoretical value. At the upper end (.3), however, the value is .2 instead of .1 as it should be. This may show the effects of guessing on the decision process. The +.8 and +1 indifference regions again yield better error probabilities than would be expected from the theory.

The ASN function for the three-parameter model (Figure 6) also shows similar results to those obtained from the one-parameter model. The +.3 indifference region required the greatest number of items,

1057

FIGURE 3
ONE-PARAMETER ASN FUNCTIONS
FOR THREE INDIFFERENCE REGIONS

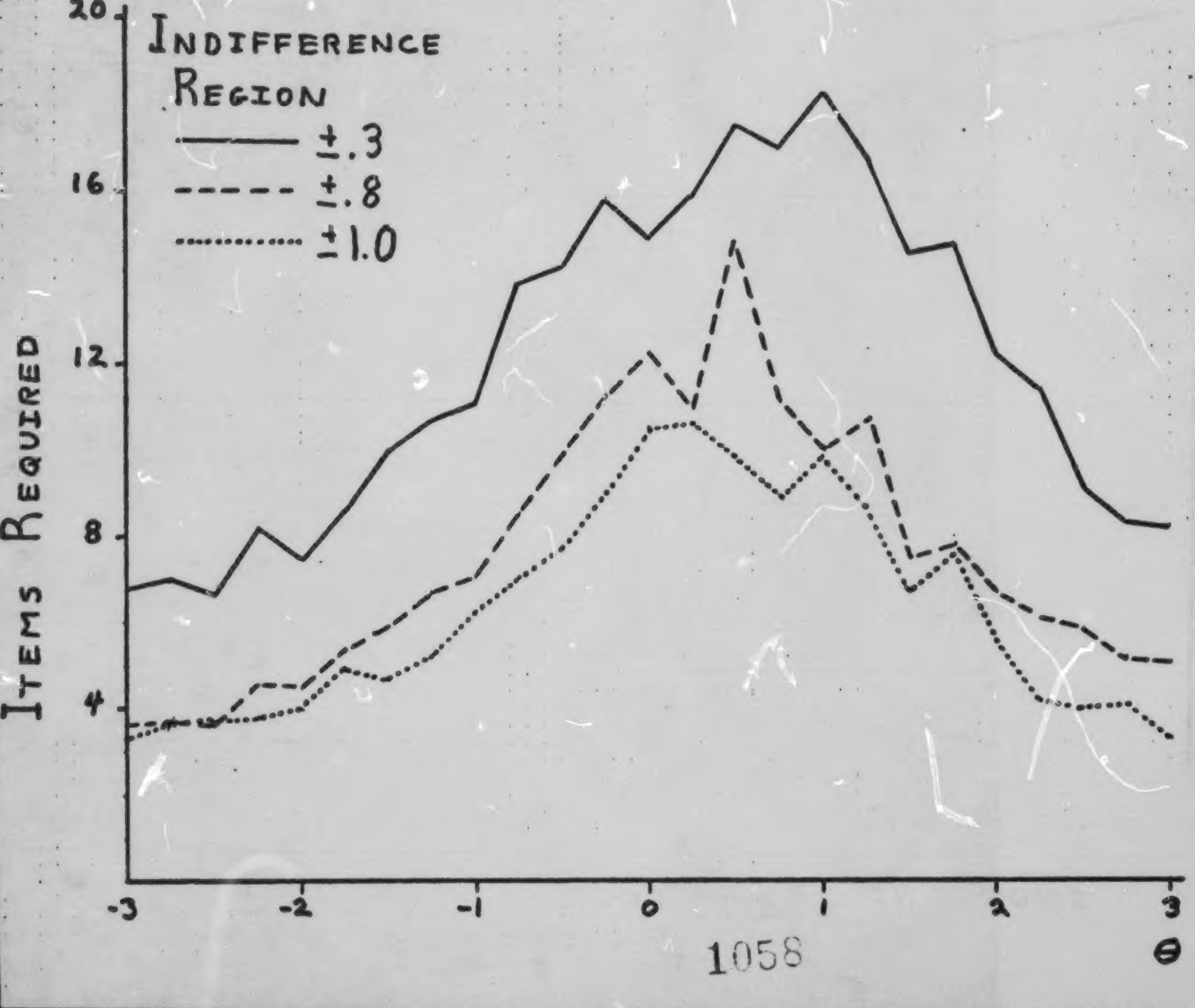
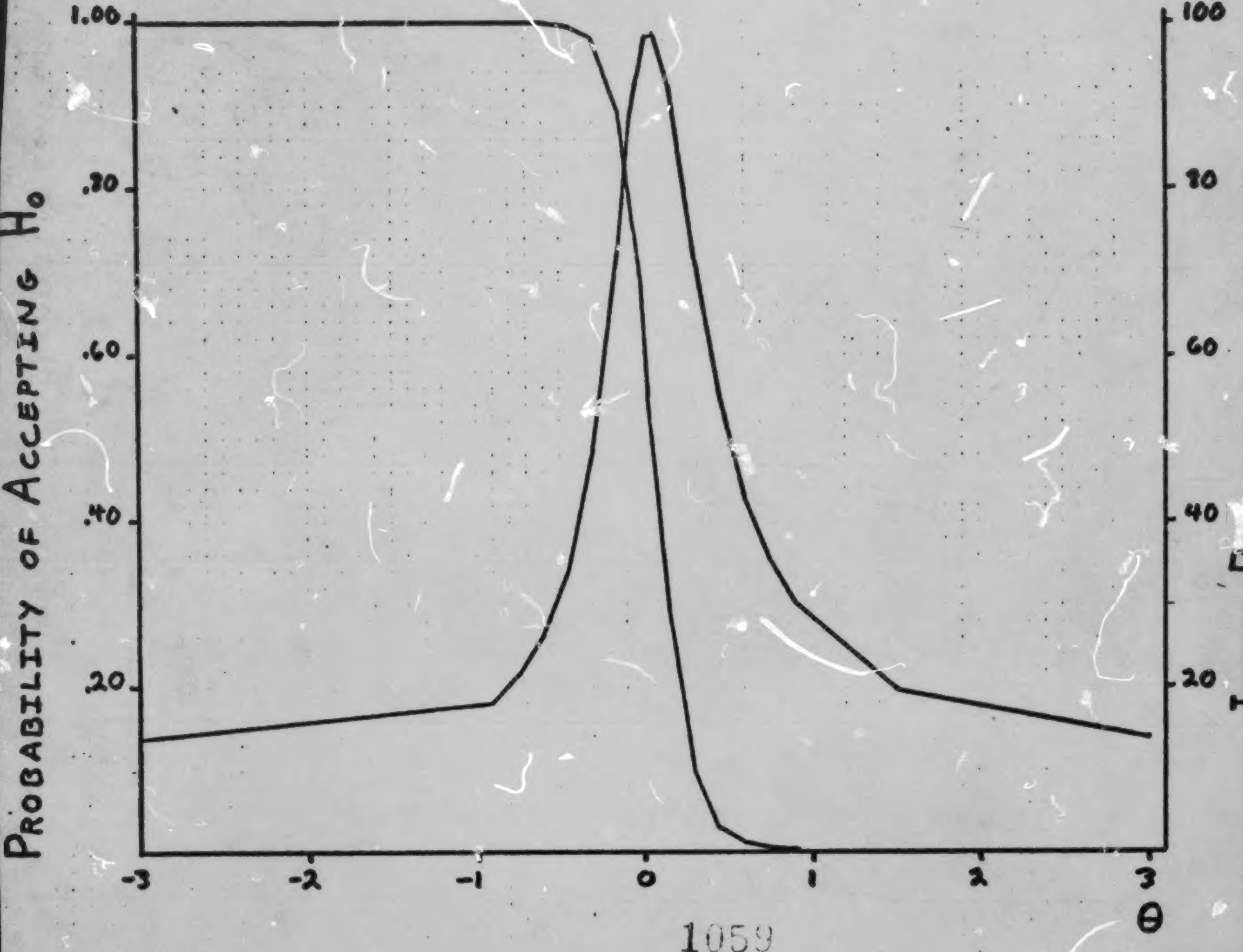


FIGURE 4

THEORETICAL OC AND ASN FUNCTIONS



while $\pm.8$ and ±1.0 require about the same. As before, the largest number was required near the decision point. However, with the three-parameter model, far less items on the average were required to make a decision. Of special note is the ASN value of about one in the -1 to -3 range on the ability scale. Decisions seem to be possible with very few items in that range.

Insert Figures 5 and 6 about here

Because of the guessing component of the three-parameter logistic model, the ASN function tended to yield more non-symmetric results than the one-parameter model. More items were required when classifying high than for classifying low to compensate for the non-zero probability of a correct response. Also, the ASN curve for $\pm.3$ indifference region was much more peaked than its one-parameter counterpart. If the simulated curves for the three-parameter model are compared to the theoretical curves presented in Figure 4, the OC functions can be seen to match the theoretical functions fairly closely while the ASN functions show that substantially fewer items were required. Over much of the ability range, as many as ten times as many items were specified by the theoretical ASN curve when unlimited identical items were assumed.

Summary and Conclusions

In the research presented here, a version of the sequential probability ratio test modified to operate within a tailored testing system has been evaluated using simulation methods. A certain amount of realism was attempted in the simulation by using latent trait item parameters derived from the calibration of a real pool of vocabulary items. Also, the simulation carried out the tailored testing within the limitations of a finite item pool and a twenty-item maximum that was imposed in an actual testing setting (Koch and Reckase, 1978).

Using the simulation data derived under these circumstances, two functions were estimated, based on either the one- or three-parameter logistic models, that can be used to evaluate the quality of the SPRT for decision making under tailored testing. The two functions are the OC and ASN functions.

Analysis of the OC functions obtained from the simulations, using several different indifference regions, yields three important results. First, the curves are very similar across the various indifference regions. This probably indicates that not enough items were available for the SPRT to function properly with the $\pm.3$ indifference region to take advantage of its theoretically greater accuracy. It should be recalled that at most twenty items were administered, and often less than that number was used because appropriate items were not available in the item pool. The three parameter OC curves were slightly steeper

FIGURE 5
THREE-PARAMETER OC FUNCTIONS
FOR THREE INDIFFERENCE REGIONS

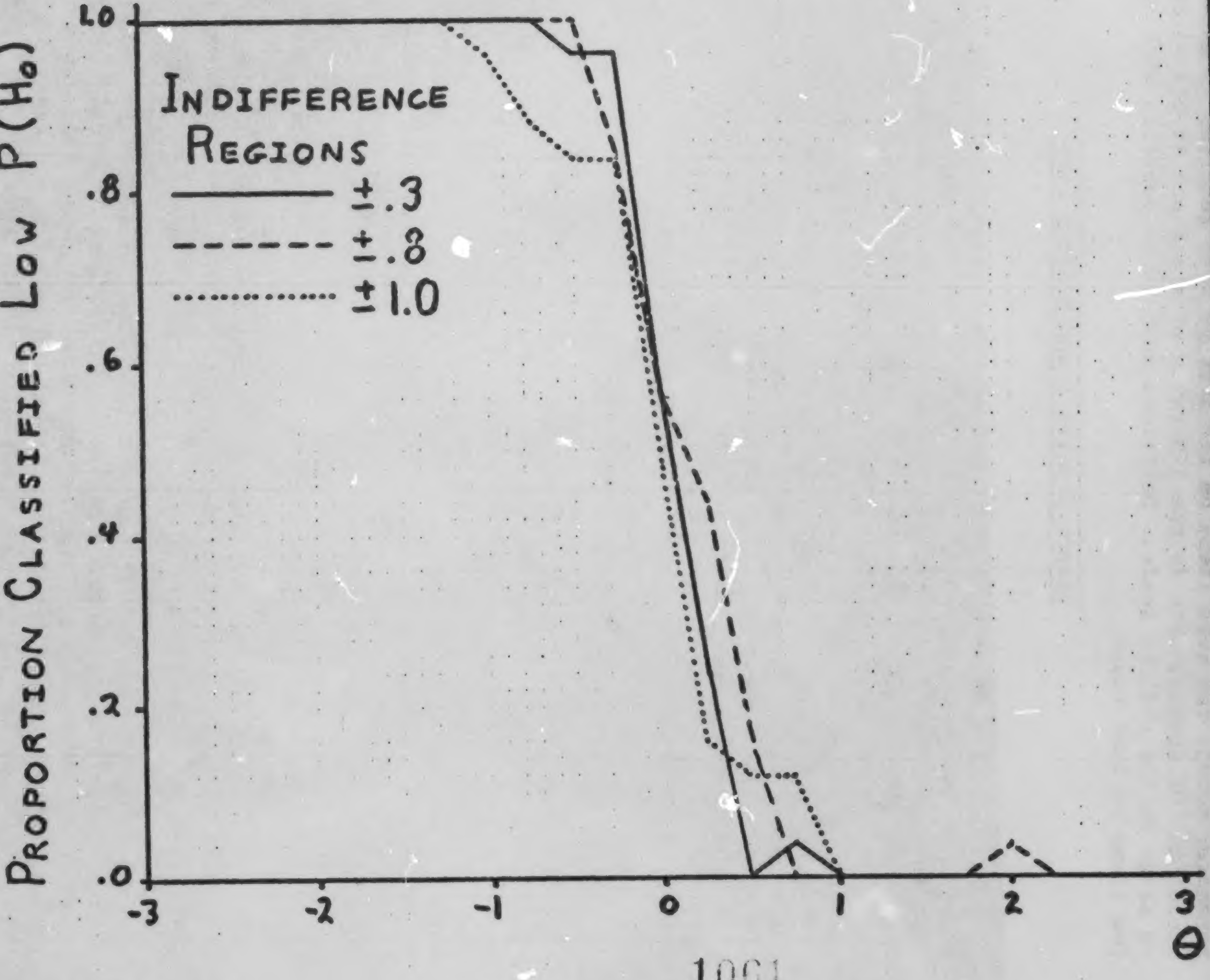
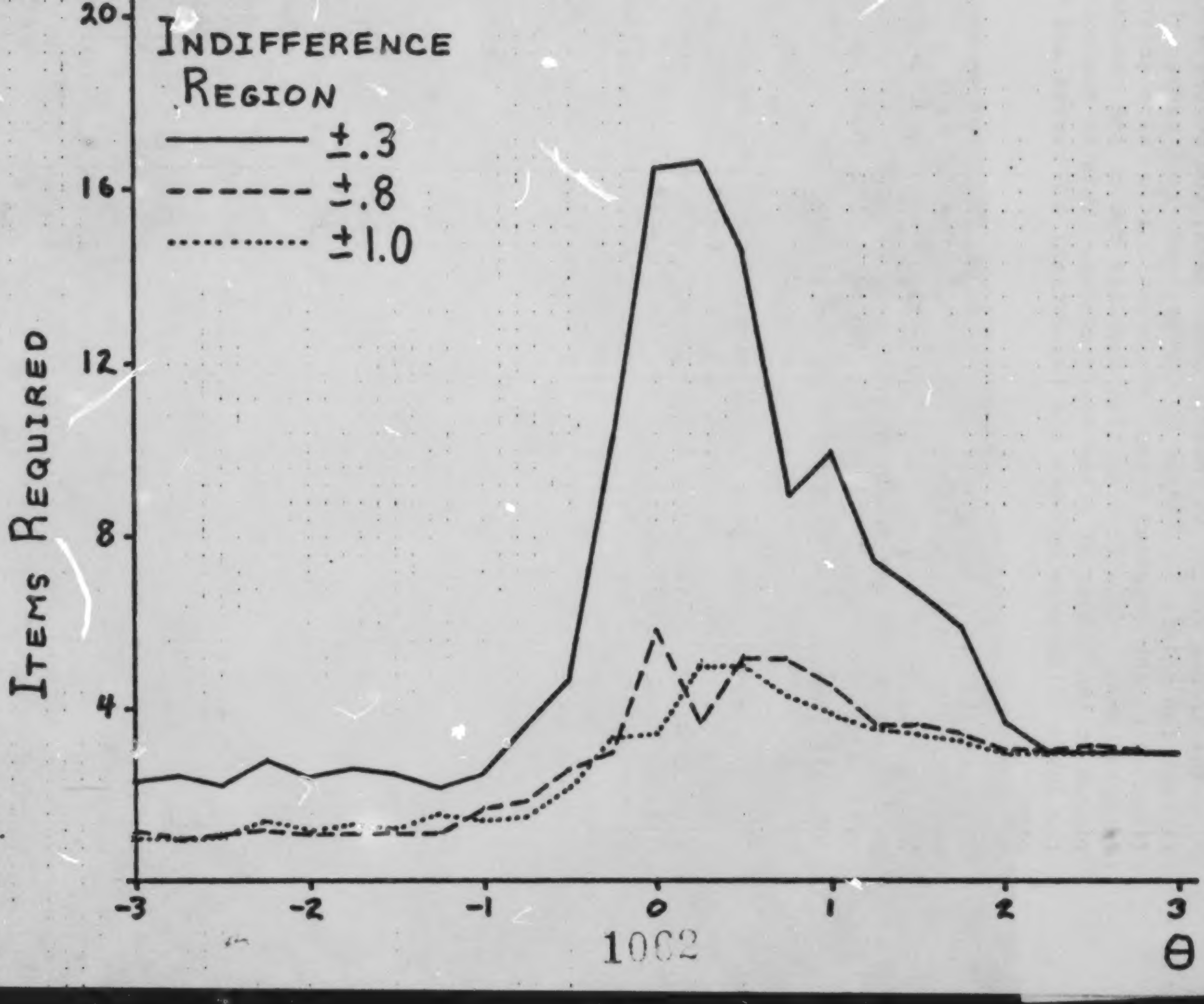


FIGURE 6
THREE-PARAMETER ASN FUNCTIONS FOR THREE INDIFFERENCE REGIONS



in the middle range illustrating the advantage of being able to select the most discriminating items with that model.

The second result is that in some cases, the curves did not pass through the points determined by the pre-set error rates. For some of these cases the obtained errors of classification were greater than the expected ones. This fact is also probably due to the restrictions placed on the number of items administered. This is demonstrated by the large difference between the theoretical ASN curves and the actual ones.

The third result of interest dealing with OC curves is that the curves at the limits of the ±.8 and ±1 indifference region tend to give better results than expected from the theoretical model. This is probably due to the advantages accrued by selecting items using the tailored testing algorithm rather than selecting them randomly from the item pool. Obviously, more research is needed to confirm these conjectures.

Directly related to the results obtained based on the OC curves are those obtained using the ASN curves. Although the OC curves were similar across indifference regions, the ASN functions show substantial differences in the number of items administered. This fact implies that the size of the indifference region should be determined by the limits imposed by the quality of the item pool and the length of the testing session. Wider indifference regions reduce the number of items required without too much loss of precision in the cases analyzed here.

Also of note, when comparing the ASN functions, is the substantial reduction in the level of these functions when proceeding from the theoretical curve, to the one-parameter curves, to the three-parameter curves. This reduction is attributed to the advantages of rationally selecting items as opposed to randomly selecting them. Since the three-parameter model has more information to use for selection, fewer items are needed to reach a decision. This is probably the most positive finding of this research for criterion-referenced measurement.

Two general conclusions can be drawn from these results. First, the SPRT has been shown to work reasonably well using a tailored testing model. Some loss of precision is present due to the stopping rules used, but the procedure seems viable. Second, the SPRT when used with tailored testing has been shown to classify relative to a cutting score with amazingly few items. Of course this finding is based on simulation results rather than live testing, but the promise of efficient and accurate classifications lends impetus for future research. Certainly these findings should be checked with live subjects to determine if the results are transferable to practical settings. However, based on the information presented here, the combination of tailored testing and the sequential probability ratio test should be considered as promising techniques for decision making in criterion-referenced testing.

References

Epstein, K. Applications of sequential testing procedures to performance testing. Proceedings of the 1977 Computerized Adaptive Testing Conference. University of Minnesota, Minneapolis, Minn. July, 1978.

Govindarajulu, Z. Sequential statistical procedures. New York: Academic Press, 1975.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. B. Criterion-referenced testing and measurement: a review of technical issues and development. Review of Educational Research, 1978, 48(1), 1-48.

Koch, W. R. and Reckase, M. D. A live tailored testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). University of Missouri, Columbia, MO: Tailored Testing Research Laboratory, June, 1978.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.; Addison-Wesley, 1968.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: current practices. Berkely, Calif.: McCutchin, 1974.

Sixtl, F. Statistical foundation for a fully automated examiner. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 1974, 6(1), 28-38.

Wald, A. Sequential Analysis. New York: Wiley, 1947.

A METHODOLOGY TO EVALUATE THE APTITUDE REQUIREMENTS OF AIR FORCE JOBS

By
Lloyd Burtch
Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, Texas

I. INTRODUCTION

Aptitude requirements for entry into various Air Force career ladders are presently determined in part by the judgement of responsible personnel and in part by tradition or precedent. A precise correspondence between the aptitude scores of Air Force personnel and the aptitude requirements of Air Force jobs is extremely important since the Air Force recruits a fixed amount of talent every year and there is more demand for this talent than one might expect. There exists an additional requirement for contingency plans should the talent pool shrink or offer fewer highly talented individuals. If such shortages were to occur, which specialties could tolerate lower aptitude requirements? Which specialties could be shredded out into different job types some requiring high level talent and some low level talent? Cost effectiveness enters the picture also. Even assuming the current talent remains unchanged, it may be more cost effective to shred some specialties into jobs with varying aptitude requirements because of differences in the actual tasks performed.

More precise information about aptitude requirements will have many repercussions for the Air Force personnel system, including procurement and training. A decision to lower the aptitude entry level for a given specialty could have devastating effects on the attrition rate for the corresponding training course if no change is made in the course curriculum. For example, if an electronics course was designed for personnel with an Armed Services Vocational Aptitude Battery (ASVAB) score of E-80 or better, the existing training program is very likely to be too difficult for those with lower aptitudes. However, the aptitude level required to be successful in the training course may or may not be the same level required for success in learning how to perform the job. It is consequently possible for the Air Force to waste talent by assigning high aptitude personnel to specialties that do not require high aptitudes; and to frustrate Air Force personnel by assigning them to jobs that do not fully utilize their talents while simultaneously neglecting other specialties in which talent is urgently needed.

The Air Force Human Resources Laboratory (AFHRL) has initiated the first systematic study to fully evaluate the aptitude requirements of Air Force specialties. The approach, originated by Dr. Raymond E. Christal, uses measures of learning difficulty at the task level to infer aptitude. The methodology was developed in an evolutionary manner from research documented by Fugill (1972, 1973). Christal (1973) as well as Maginnis, Uchima and Smith (1975) have further described this technology. The present paper will describe the development of task difficulty benchmark scales, their application, and will include a brief discussion of the results.

II. BASIC CONCEPTS

Task Difficulty

Task difficulty was operationally defined in terms of the time it takes to learn to perform a task satisfactorily. Based on Fugill's demonstration (1972) of high relationships between task difficulty and task aptitude ($r>.89$), this research has been conducted under the assumption that the aptitude level required to learn a job can be inferred from task difficulty, as defined above, of the tasks that make up the job.

Benchmark Scales

A technique was required that would allow for the comparison of the learning difficulty of tasks both within and across Air Force specialties. A difficulty scale, using one or more tasks at each scale point as examples of that level of difficulty, would fill this need. Table 1 presents a simple example of such a scale. Task-anchored or benchmark scales were demonstrated to produce more reliable ratings of several task factors than did numerically anchored scales in a study by Peters and McCormick (1966). The feasibility of using task difficulty benchmark scales has been demonstrated by Fugill (1972, 1973).

Table 1. Example Benchmark Scale

Level 1 - Very Low Task Difficulty
    Visually inspect batteries

Level 2 - Low Task Difficulty
    Check fuse indication

Level 3 - Average Task Difficulty
    Adjust transmissometer projector lamp voltages

Level 4 - High Task Difficulty
    Trouble-shoot wind measuring sets

Level 5 - Very High Task Difficulty
    Trouble-shoot aircraft flight control circuits

Aptitude Areas

There are four aptitude areas in the Air Force personnel testing system: general, administrative, mechanical and electronics. This research does not question the appropriateness of these areas; it is concerned with the relative order of aptitude area score requirements for specialties and jobs within each of those areas.

III. DEVELOPMENT OF BENCHMARK SCALES

Task difficulty benchmark scales have already been developed for the electronic, mechanical and general aptitude areas. The approach was similar for all scales, but the mechanical scale will be used as an example.

A general description of the scale development effort was presented by Hart (1977) at last year's Military Testing Association Conference in San Antonio. The 15 specialties shown in Table 2 were selected for the mechanical scale development. These specialties are representative both of the complexity and the variety of tasks within the mechanical aptitude area.

Table 2. Mechanical Specialties
(N Task and ASVAB Cut Off)

Air Force Specialty	N Task	Mech ASVAB Cut Off
464X0-Explosive Ordnance Disposal Spec.	551	60
431X0-Helicopter Mech.	577	50
542X2-Electrical Power Production Spec.	592	50
546X0-Liquid Fuel Systems Mech.	1018	50
427X1-Corrosion Control Spec.	457	50
361X0-Outside Wire and Antenna Mech.	476	40
423X2-Aircrew Egress Systems Mech.	376	40
423X3-Aircraft Fuel Systems Mech.	297	40
426X2-Jet Engine Mech.	415	40
552X0-Carpenter	563	40
552X5-Plumber	407	40
566X1-Environmental Support Spec.	556	40
551X1-Construction Equip. Operator	927	40
427X3-Fabrication and Parachute Spec.	553	40
551X0-Pavements Maint. Spec.	927	40

Table 3. Estimates of Interrater Reliability

	Specialty	N (Rater)	R_{KK}
464X0	Explosive Ordnance Disposal Spec.	88	.96
431X0	Helicopter Mech.	100	.97
542X2	Electrical Power Production Spec.	58	.96
546X0	Liquid Fuel Systems Mech.	81	.96
427X1	Corrosion Control Spec.	43	.88
361X0	Outside Wire and Antenna Mech.	38	.93
423X2	Aircrew Egress Systems Mech.	53	.88
423X3	Aircraft Fuel Systems Mech.	26	.93
426X2	Jet Engine Mech.	83	.94
552X0	Carpenter	68	.93
552X5	Plumber	116	.97
566X1	Environmental Support Spec.	56	.94
551X1	Construction Equip. Operator	83	.97
427X3	Fabrication and Parachute Spec.	73	.94
551X0	Pavements Maint. Spec.	72	.97

Relative ratings of task difficulty are routinely obtained in conjunction with job inventories and occupational surveys conducted by the USAF Occupational Measurement Center, Lackland AFB. These data, obtained from incumbent supervisors, are collected on all tasks in the job inventories and are provided to AFHRL for research purposes. Table 3 reflects the estimates of interrater reliability (Lindquist, 1953) and the number of raters for the 15 mechanical specialties. Using these data and the criteria outlined in Table 4, forty tasks were selected from each specialty to establish a set of 600 benchmark tasks.

Table 4. Task Selection Criteria

1. Eliminate supervisory tasks
2. Capture range of difficulty
3. Select on High Rater Agreement (Low SD)
4. Tasks performed by first termers
5. Prefer well known tasks
6. Prefer easily observed tasks
7. Face validity

In preparation for selecting the tasks from the benchmark set to represent the 25 points on the benchmark scale, a panel of mechanical experts, provided by an Air Force contractor, was asked to provide a rank-ordering of the 600 tasks. Each panel member, after accumulating detailed information on each task, provided an independent rank-order of the set of 600 tasks. The task requiring the least learning time was assigned number 1 and the task requiring the greatest learning time was assigned number 600. The estimate of interrater reliability was very high (R_{KK}=.97, N=8). This result demonstrates that a panel of work area experts can work within our definition of task difficulty, collect detailed information in the field at the task level, and provide highly reliable rank orderings of a large number of tasks selected from a given specific work area.

To address the matter of validity, the contractor's ranking data were correlated with the field supervisor's relative ratings referred to earlier. These correlations were computed using mean ranks and ratings on the forty tasks from each of 15 specialties separately; results are summarized in Table 5. These coefficients provide some substantiation of the validity of the data collection procedure, the definition of learning difficulty, and of the data itself.

Table 5. Correlations between Mean Ranks and Mean Ratings of Forty Tasks

Specialty	r
464X0	.87
431X0	.91
542X2	.87
546X0	.85
427X1	.81
361X0	.77
423X2	.83
423X3	.79
426X2	.74
552X0	.76
552X5	.57
566X1	.76
551X1	.82
427X3	.81
551X0	.73

Benchmark Task Selection

Two tasks were selected to represent each of the learning difficulty levels of the 25-point scale. A systematic procedure was developed to insure that the selected tasks represented the distribution of the mean ranks of the 600 tasks. In addition, the criteria summarized in Table 4 were again applied as appropriate. Face validity was even more important in this task selection process than it was in the prior process in as much as the tasks were to be used as examples that would anchor the various points on the scale. That is, the tasks on the mechanical scale must appear to be mechanical tasks to the extent possible.

A sample of the 50 selected tasks (two for each of 25 points) along with mean and standard deviation from the ranking process is at Table 6. The mean standard deviation for all 600 tasks was 62.8. Table 6 indicates the type of tasks selected as well as the relatively high rater agreement for most of them.

Table 6. Example Benchmark Tasks - Mechanical Scale

Level	Task Title	$\bar{X}$	SD
1	Police Grounds for Litter	1.50	.87
1	Police Open Storage Areas	3.50	1.73
5	Clean Life Preservers	26.38	13.77
5	Dig Ditches by Hand	27.00	14.41
10	Clean or Regap Spark Plugs	136.38	53.97
10	Caulk Areas Around Windows, Sinks or Bathtubs	140.63	105.52
13	Install or Replace Water Fountains	307.38	77.31
13	Disassemble or Clean Conventional Fuel Gate Valves	306.13	83.64
15	Perform Preoperational Inspections of Engine after Engine has been on long Standby	401.63	88.07
15	Install or Replace Formica on Counter-tops or Splashboards	404.13	74.44
20	Install Tail Roter Assemblies on Helicopter Aircraft	562.50	24.09
20	Read and Interpret Schematic or Wiring Diagrams	562.00	58.41
25	Troubleshoot Installed Engines	599.38	1.32
25	Troubleshoot Systems for Breaker Trip-Outs	595.38	5.20

IV. PROCEDURAL GUIDE

Accurate application of the benchmark scale requires detailed knowledge of both the task being rated and the reference tasks at each level of the scale. A procedural guide has been assembled for each scale describing the reference tasks. This guide is for the use of the panel of expert raters who will actually apply the scales. There are two parts: Part I introduces each panel member to the task of assessing learning difficulty and rating the tasks; Part II presents the 25-point scale and provides a one page description of each of the 50 tasks on the scale. This description includes the level of the task on the scale, the title of the task, the specialty from which it was selected, a narrative on any specific equipment associated with the task, a narrative describing the actual task performance, and an explanation of the skill and knowledge required to learn the task. Examples of these descriptions, taken directly from the Mechanical Procedural Guide (Hart and Pulliam, Note 1), are at Figures 1 and 2.

Figure 1. Level 10 Task Description

Level 10: CLEAN AND REGAP SPARK PLUGS (Electrical Power Production Specialist - AFSC 54350

Equipment: The task concerns gasoline engines of one or two cylinders, driving service equipment such as air compressors. These engines are part of the support equipment in an electrical power generating station.

Task Description: The task requires standard hand tools and an air blast powered spark plug cleaner which blows an abrasive against the plug base to clean insulator and electrodes. Work is performed in the power station. The mechanic removes plugs from the engine, using a socket wrench. He cleans the plug by inserting it into a hole on the cleaning machine, and pressing a valve to release a blast of abrasive against the plug base. After a few seconds he removes the plug, inspects it visually for clean ceramic, and (on some machines) inserts it in a second hold for a pressure test. Defective plugs are thrown away. He then checks the gap using a gap gauge (with feeler wires), and corrects any error by bending the outer electrode inward, using a slotted wrench which is often part of the gap-gauge handle. He puts a new plug gasket on the plug and torques the plug back in place.

Skill/Knowledge Required: The task requires knowledge of standard hand tools, including a torque wrench. Since there is likely to be no T.O. for the engine concerned, the mechanic must know the general procedure for cleaning and gapping a plug, and that 25 foot pounds is the usual plug torque. Airmen who qualify for entry into this field usually have some knowledge of this task before their enlistment.

1071

Figure 2. Level 25 Task Description

Level 25: TROUBLESHOOT INSTALLED ENGINES (Jet Engine Mechanic - AFSC 42652)

Equipment: This task is performed on jet engines installed on aircraft. Troubleshooting includes isolation of failure within the engine or confirming that a failure is not in the engine but some related subsystem.

Task Performance: Troubleshooting typically begins with a pilot write-up. Interpretation of these write-ups is often difficult. The isolation process depends upon the failure sympton observed. Oil leaks, which are the most common problems require that all oil be cleaned from the exterior of the engine, the engine and oil systems are isolated by attaching vibration sensors at different locations around the engine and then running the engine to look for abnormal vibration sources. Other problems such as fuel leaks, throttle rigging, fuel control, and electrical problems require coordination with other subsystem specialties to isolate the problem between the engine and related systems.

Skill/Knowledge Required: Learning troubleshooting is accomplished by exposure and is not formalized. It requires:

(a) A complete knowledge of engine operation and its interface with related aircraft subsystems.

(b) Ability to use and understand the readings of pressure gauges, vibration sensors, and heat gauges.

(c) That the mechanic be cockpit qualified to enable him to run up the engine.

(d) An ability to read and interpret the appropriate Technical Orders.

(e) Coordination with the efforts of other subsystem specialists to isolate problems in the interaction of the engine and related aircraft systems.

It is mandatory for each rater to fully absorb the contents of the guide prior to using the scale. Part 1 of the guide calls for a practice period of actual study and application prior to operational use of the scale.

V. APPLICATION OF BENCHMARK SCALES

The intention is to ultimately apply the scales to all available enlisted specialties in the Air Force. Data collection and analysis is underway. Because analysis is not complete, information to finalize the evaluation of the aptitude requirements in specific specialties is not yet available. Presented here is a brief discussion on how the method is to be applied.

Typically 60-70 tasks are selected from each specialty to be evaluated. These tasks will be selected using criteria similar to those used in selection of the benchmark set. The tasks will be individually studied in depth at both the technical school and at two or more operational work sites. A typical panel will be made up of 12 members with two teams of six visiting separate locations. After accumulating as much data as feasible on each task, the panel members will independently provide 1-25 point ratings of learning difficulty for all 60-70 tasks in each specialty. These ratings (for a sample of tasks within each specialty) can be used to estimate the learning difficulty of all tasks in a specialty using traditional statistical procedures for estimations.

VI. DATA ANALYSIS

The Comprehensive Occupational Data Analysis Program (CODAP) package developed by AFHRL is the data analytic tool being used in the analysis of these data. The CODAP system is ideally suited for this job. Programs are readily available to provide all necessary analysis for the project.

The contractor's benchmark ratings and the supervisor's relative ratings of the same 60 tasks are input to a two variable multiple regression problem for each specialty. The resulting equation is then to be applied to the supervisor's relative ratings of all tasks in the specialty. This process will result in the prediction of a 1-25 point rating mean for each task in the specialty. These predicted difficulty levels are, in turn, used as input to the CODAP system for the computation of average task difficulty for a variety of groups, and job types within each occupation. For example, the average task difficulty for first term airmen will be computed for each specialty and will be comparable across all specialties in an aptitude area. Similar computations will be made on other combinations of tasks and/or job incumbents.

VII. PRELIMINARY RESULTS

The analysis completed to date has resulted in demonstration of the efficacy of the method. Interrater agreement estimates with 12 raters rating 60 tasks from each specialty have ranged from .88 to .98. These results have convinced us that the scale, in hand with the procedural guide, can be reliably applied by knowledgeable work experts.

Some preliminary correlational analysis has been completed with positive results. Correlations between the two teams of raters have ranged from .82 to .94. Correlations between the ratings of relative difficulty and the benchmark ratings are ranging from .71 to .94. Both of these results are indicative of the validity of our methodology. Further data collection and analysis will be much more conclusive. An illustration of the planned format of the data is provided in Figure 3.

Typically 60-70 tasks are selected from each specialty to be evaluated. These tasks will be selected using criteria similar to that used in selection of the benchmark set. The tasks will be individually studied in depth at both the technical school and at two or more operational work sites. A typical panel will be made up of 12 members with two teams of six visiting separate locations. After accumulating as much data as feasible on each task, the panel members will independently provide 1-25 point ratings of learning difficulty for all 60-70 tasks in each specialty. These ratings (for a sample of tasks within each specialty) can be used to estimate the learning difficulty of all tasks in a specialty using traditional statistical procedures for estimations.

VI. DATA ANALYSIS

The Comprehensive Occupational Data Analysis Program (CODAP) package developed by AFHRL is the data analytic tool being used in the analysis of these data. The CODAP system is ideally suited for this job. Programs are readily available to provide all necessary analysis for the project.

The contractor's benchmark ratings and the supervisor's relative ratings of the same 60 tasks are input to a two variable multiple regression problem for each specialty. The resulting equation is then to be applied to the supervisor's relative ratings of all tasks in the specialty. This process will result in the prediction of a 1-25 point rating mean for each task in the specialty. These predicted difficulty levels are, in turn, used as input to the CODAP system for the computation of average task difficulty for a variety of groups, and job types within each occupation. For example, the average task difficulty for first term airmen will be computed for each specialty and will be comparable across all specialties in an aptitude area. Similar computations will be made on other combinations of tasks and/or job incumbents.

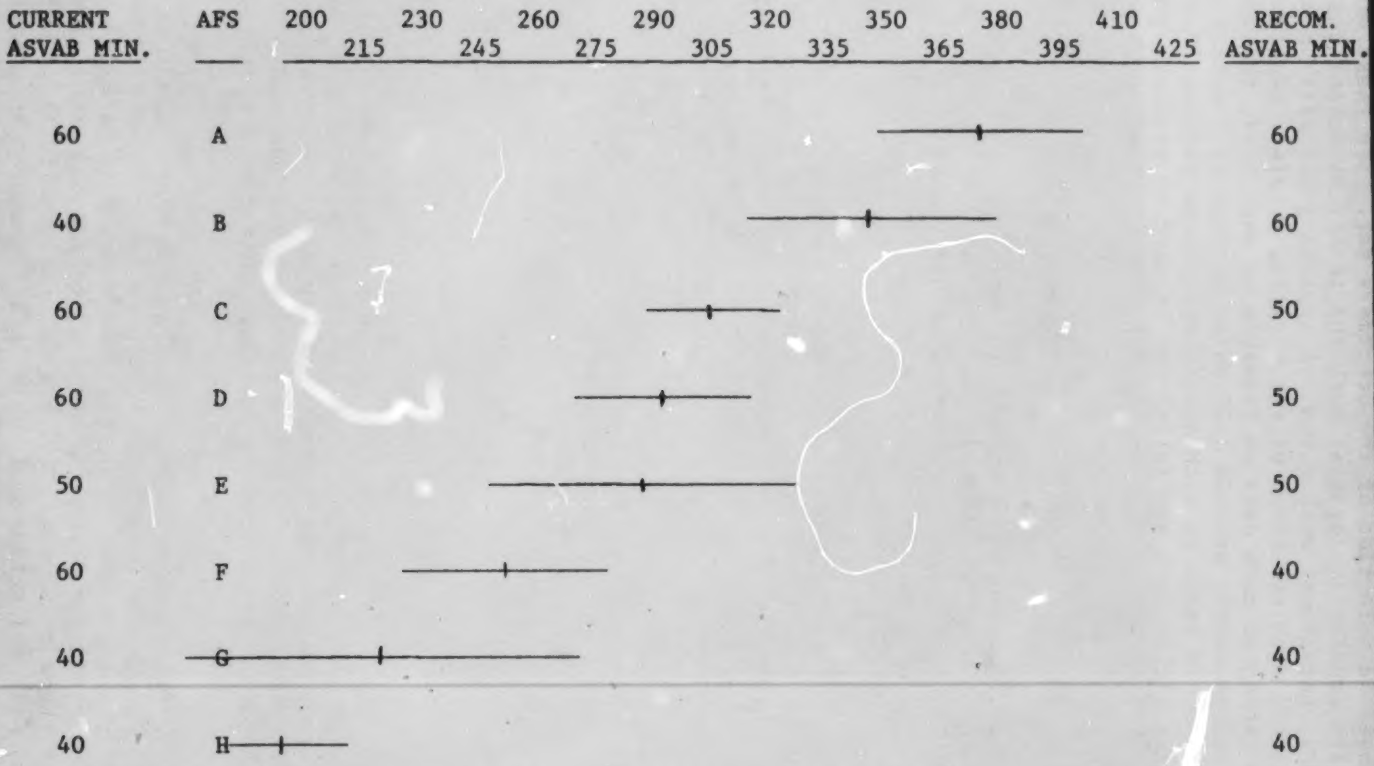
VII. PRELIMINARY RESULTS

The analysis completed to date has resulted in demonstration of the efficacy of the method. Interrater agreement estimates with 12 raters rating 60 tasks from each specialty have ranged from .88 to .98. These results have convinced us that the scale, in hand with the procedural guide, can be reliably applied by knowledgeable work experts.

Some preliminary correlational analysis has been completed with positive results. Correlations between the two teams of raters have ranged from .82 to .94. Correlations between the ratings of relative difficulty and the benchmark ratings are ranging from .71 to .94. Both of these results are indicative of the validity of our methodology. Further data collection and analysis will be much more conclusive. An illustration of the planned format of the data is provided in Figure 3.

Figure 3. Relative Aptitude Requirements for First Term Jobs In 8 Specialties (Hypothetical Data)

Relative Difficulty (Bar = ± 1SD)



A brief comparison of the column containing current ASVAB minimum with the column reflecting recommended ASVAB minimum indicates that there is evidence of misalignment of the aptitude requirements of these eight specialties. Specifically, Figure 3 indicates that some specialties may have a high current minimum aptitude requirement but may actually have a much lower required minimum (e.g., specialties C, D and F). The opposite is true for specialty B. Other specialties will be found to cover an extremely wide range of jobs (indicated by the length of the horizontal lines on Figure 3) suggesting that the specialty itself might be shredded out in some fashion . The information contained in Figure 3 is not based on actual data; but data of this type will soon be available on approximately 200 specialties. Changes in aptitude requirements require a total systems approach, and we do not intend to release any data in a piece-meal fashion.

VIII. CONCLUSIONS AND FUTURE PLANS

The analysis of data to date indicates that we have developed a methodology which will enable us to evaluate aptitude requirements at the task, job, and occupation level. The benchmark scale approach results in the collection of difficulty data at the task level that is comparable across all tasks within an aptitude area regardless of specialty. The results of the data analysis to date are sufficient to conclude that the total technology is based on a sound approach and analysis methodology.

There are studies in process that address the matter of longevity of the data; that is, how long will these data reflect the requirements of the specialty. Preliminary results indicate that the contractor benchmark data may be useful in assessing the learning difficulty of the specialty for several years. The difficulty scale is anchored with tasks that should not easily become obsolete because of the task selection process. First, to the extent possible, tasks were selected that were well known to mechanical workers; and second, extreme care was used in documenting each task in the procedural guides. Primarily for these reasons, it is not necessary for the tasks on the scales to even remain in active occupational task inventories to be effective. The scale will remain an effective tool as long as experts in the work area can comprehend the terminology used and the written documentation provided in the procedural guide. Not only will the scale and the benchmark data be useful in years to come, but the scales as they are will also be useful in examining the difficulty level of future tasks as they are added to job inventories. This procedure will allow the evaluation of the aptitude requirements of new specialties and/or tasks as they become a part of Air Force work.

Implementation of the results of this project is anticipated in FY 80 or 81. The primary procedure for implementation is to change the aptitude minimums as listed in AF Regulation 39-1. The results will also be implemented through the computerized job-offer system used by the AF Recruiting Service. Plans for this form of implementation are currently being prepared. We also plan to develop a total implementation package that will include complete impact analyses with recommendations *for coordinated* changes in the length and difficulty of Air Force resident school training courses.

There are three significant areas where cost avoidance should be achieved as a result of this research. Contingency plans for talent shortages will be available as a product of this effort. These plans will enable the Air Force to specifically plan for talent shortages in any specific specialty or across all specialties. Another product will be a more defensible position for aptitude requirements in the case of court actions. The present system, which excludes many individuals from entering Air Force jobs based on a "cut-off aptitude score," has no objective data to support its use. This research will provide data on the learning load requirements for each job. Another product will be an improved match-up of Air Force talent and job requirements. Improving this match of talent with requirements can have effects on job attitude, retention, recruiting, and training, to name just a few.

REFERENCES

Christal, R. E. *The United States Air Force Occupational Research Project*. AFHRL-TR-73-75. AD-774 574. Lackland AFB, TX.: Occupational Research Division, Air Force Human Resources Laboratory (AFSC), January 1974.

Fugill, J. W. K., SqdnLdr, USAF (RAAF) *Task difficulty and task aptitude benchmark scales I. Mechanical and electronic career fields*. AFHRL-TR-72-40. AD-754 848. Lackland AFB, TX.: Personnel Research Division, Air Force Human Resources Laboratory (AFSC), April 1972.

Fugill, J. W. K., SqdnLdr, USAF (RAAF) *Task difficulty and task aptitude benchmark scales for the Administrative and General career fields*. AFHRL-TR-73-13. AD-771 677. Lackland AFB, TX.: Personnel Research Division, Air Force Human Resources Laboratory (AFSC), October 1973.

Hart, F. L. Study of task difficulty using field terms and the AFHRL task anchored scales. *Proceedings of the 19th Annual Conference of the Military Testing Association*. San Antonio, Texas, 832-839, October, 1977.

Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Hougton Mifflin Co., 1953, 359-361.

Maginnis, E. B., Uchima, A., and Smith, C. E. *Establishing Aptitude Requirements for Air Force Jobs*. AFHRL-TR-75-44 (Vol I, II & III). Vol I, AD-A023 250; Vol II, AD-A022 206; Vol III, AD-A022 250. Occupational and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), October 1975.

Peters, D. L. and McCormick, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology*, 1966, 50(1), 92-96.

REFERENCE NOTE

1. Hart, F. L. and Pulliam, R. *Procedural guide for use of mechanical benchmark scale*. Unpublished manuscript, 1977. (Available from Air Force Human Resources Laboratory, Brooks AFB, TX 78235).

SECTION 12

EXAMINATION ITEMS

OBJECTIVE EVALUATION OF CORRESPONDENCE COURSE ITEMS

Andrew N. Dow, Ed.D.

Naval Education and Training Program Development Center
Pensacola, Florida 32509

Every word that is spoken or written is evaluated to some extent by someone. Those of us who prepare training materials feel more at ease when we get evaluative feedback from the performance of our materials. It is in reply to such quests for feedback that this paper has been prepared. However, before we go further, we must define and describe our subject matter.

In this presentation, "correspondence course" refers to the series of interrogatories that accompanies the text. The individual interrogatories are the items of the correspondence course. This material converts a book, or other text materials, into a self-teaching course.

Most of the items which comprise the correspondence course, bear a strong resemblance to multiple-choice test items; some ask a question that is followed by several possible answers, while others contain a stem that is an incomplete statement followed by several possible completions. In spite of the superficial resemblance to the typical objective test item, the primary purpose of the correspondence course item is instruction. Evaluation, which is the primary purpose of the test item, becomes the secondary purpose of the course item. Conversely, instruction, which is the primary purpose of the course item, is the secondary purpose of the test item. It must also be recognized that some course items are more evaluative than others, while some are almost pure teaching items, too easy to have any evaluative function.

Regardless of their function, course items need to be evaluated. An item that is unrelated to the course and its learning objectives is a waste of paper and the time of the student. Further, that item may be occupying the space of an as yet unwritten effective item. Like any other training material, the items of a correspondence course can be evaluated when they are reviewed by knowledgeable persons. Optimally, the person who originally prepares the items takes a second look some time after preparing them, and a co-worker also gives them a critical review. This constitutes internal review. External review consists of critical evaluation of one or more items by an uninvolved person.

Evaluation by review has a number of shortcomings. The most obvious is the amount of manpower required to do a good job. A single review is time consuming; multiple reviews are more so. Since there is a good chance that any review will be biased, several reviews may overcome the bias if the reviewers hold relatively diverse viewpoints. When there are several reviewers of diverse viewpoints, in addition to those that are internal and involved, there are those that are external and impersonal. The internal reviewer who has been involved in the development of the materials brings a sophistication that is as necessary as the impersonality of the noninvolved external reviewer. With a diverse group of reviewers, there may be little agreement; someone still must decide which criticisms to accept or reject and must synthesize their aspects. All of this increases the manpower demands of such a process, and, even with intersubjective agreement, doubt still remains as to its objectivity and validity.

A system of item evaluation that requires somewhat less manpower is based upon the surface resemblance of the course item to the typical objective examination item. This system employs item response counts as used in test item analysis (2). These counts are objective and reliable, requiring very little manpower, but there seems to be no consensus as to the meaning of the counts, nor how they can be used to improve the courses. Thus, it is appropriate that we look at some of the possible causes of some of the several levels of correct response counts.

Some items will be answered correctly by almost everyone, e.g., giveaways--"When was the War of 1812 fought?"--which one can answer without having taken the course and without any great fund of general knowledge. This is the worst kind of high percentage correct item. The best kind of high percentage correct item is one that is well covered in the text; the text materials are comprehensive and comprehensible, and the course item is not ambiguous. These two kinds of easy items are the extremes. Other items will be answered correctly by a high percentage of the respondents because the items are based on information available to most of them--sometimes called common knowledge. If an item such as this has some bearing on the rest of the course, there may be justification for keeping it. If a common knowledge item is related only vaguely to the course subject matter, there is no reason to retain it. Sometimes, high percentages correct are the result of compromise--the word has gotten out on some of the items. This is particularly likely to occur with a popular course. If compromise is relatively universal, then a rewrite is in order; the basic material can be covered from a different viewpoint. Thus, we have four of the possible reasons for a high percentage of correct answers, and only one of them is really desirable from a pedagogical point of view.

A large number of items will be answered correctly by a moderately high percentage (60%-80%) of those taking the course. An obvious reason for this in some instances is that the text covers the material, but not as well as it does for the good items that are correctly answered by a higher percentage. Some others will fall into this group because the item is not well phrased; the text is not at fault, the item is. A third kind of item joins these others just because its subject matter does not stimulate thought and/or learning; both the text and the item are well worded and course related, but the material just is not remembered well. Ofttimes this kind of material is part of a series of building blocks and is essential to the course. Some items will fall into the moderately high percentage correct group because they are based upon general knowledge that is not universal knowledge. Others will wind up here because they are relatively difficult but have been compromised to a limited degree.

Then, there are those items that are correctly answered by only a very small percentage of the respondents. Some of these are in this group because of exaggerations of the conditions that produced the items for the moderately high percentage correct group. In addition, some of the items are not answered by many respondents because the text does not cover the material well enough for many to get the item correct. Some other items fall into this statistical group because the item is worded ambiguously, and most of the respondents choose a wrong interpretation. Also, there are some items that very few answer correctly because the item structure is such that they become high-level ability items, even though all the material needed to answer them can be found scattered about the text.

We have just looked at some of the reasons that course items are answered by a certain percentage of those who take the course. This is not an exhaustive list of the reasons behind item behavior, but it is a start. Obviously, the item count percentages are not diagnostic. Without careful analysis (subjective) there is no way to tell whether an item is adequate as it stands, needs some revision, or should be thrown out.

Some of the dilemmas raised by the simple item analysis type response count can be resolved by using quasi-experimental designs (1) which incorporate several item counts. We will first describe three possible designs and then discuss the probable outcomes of using each of them. Each of these designs involves a process analogous to pretesting; some students will go through the interrogatory items of the course, answering each before studying the text materials. These same students will take the course after being allowed to study; other groups will take the course under varying sets of conditions. These are described in the following paragraphs.

The first procedure (QE 1) is comparable to a simple "Test-Retest" design. All persons who participate will take the course, answering the items without having access to the text materials. While the course is not, strictly speaking, a test, this participation without study will be termed a "pretest." Then, these same participants will study the course text and answer the course items. This will be known as the "post test." The response counts from

1082

these two uses of the course items yields three P values (percentage of respondents answering correctly) for each item--a pretest P value to be called Pre P; a post test P value to be called Post P; and a differential P value derived by subtracting the Pre P for an item from its Post P. The differential P value will be designated Dif P. The Pre P gives an indication of how much the item depends upon common or precourse knowledge; a Pre P of 50 indicates that half of the participants were able to select the correct response without benefit of the course text. The Post P of an item is an indication of the general difficulty of the item, but does not show whether the item was answered from general, precourse knowledge or from course derived information. Dif P indicates how well the item is related to the text of the course; the larger the Dif P (in relation to the Pre P), the more the item depends upon the text material. Compromised items and items that can be answered from general knowledge will have a rather low Dif P.

The second procedure (QE 2) is designated "Test-Retest with Post Control." In addition to participants as used in the first procedure, this procedure calls for a control group. These two groups (group X, the experimental participants, and group C, the control) should be selected or matched by one of the systems recommended by Campbell and Stanley (1). Group X is handled just like the participants in the first procedure, and the data derived are of the same type. Group C takes the course in the regular fashion without a pretest; the item counts from Group C should be representative of the usual course takers. The item count from Group C is designated Post P_C, and that from the post test data from group X as Post P_X. Thus P_C and P_X can be compared for each course item.

A third procedure (QE 3) for the evaluation of course items is called the "Test-Retest with Dual Control." This procedure calls for groups X and C as in the Test-Retest with post control and, in addition, a second control group called group CC. Groups X and C participate the same way as in the Test-Retest with Post Control. Group CC takes the items twice with group X, but does NOT have access to the text for the second taking of the items. The additional data yielded by this procedure are labeled Pre P_{CC}, Post P_{CC}, and Dif P_{CC}.

Table 1 summarizes the three procedures and compares them with taking a course in normal fashion.

Each of the three procedures entails more work than a simple count of the responses of a group of course-taking students. What benefits are derived from each of these procedures? What are the limitations of the three? We shall attempt to answer these questions by examining each of the three.

Table 1

Comparison of Normal Procedure of Course Taking and Three Experimental Designs

Procedure	Group	PreTest (Items)	Training (Text)	Post Test (Items)
Normal	S	-	S	S
QE 1	X	X	X	X
QE 2	X	X	X	X
	C	-	C	C
QE 3	X	X	X	X
	C	-	C	C
	CC	CC	-	CC

Before examining the three quasi-experimental designs, we should look at the simple count of responses. The simple count indicates the percentage that responds successfully to each item after having had access to the text materials. This count does not indicate how much of the success of the respondents can be attributed to their exposure to the text, how much to knowledge that they had before they started the course, and how much to incidental learning that occurred concurrently with taking the course. We have previously pointed out some of the reasons for an item's being answered by a given percentage.

The first of the procedures (QE 1) the simple Test-Retest, yields three kinds of data: Pre P, Post P, and Dif P. The Pre P values give a good indication of the extent to which the items depend upon general knowledge that the students had before they started the course. Ideally, these values are low, about 25 or less. The Post P indicates the general level of achievement after the course is completed by students that have been primed by the pretesting. Dif P is an indication of how much the students improved during the period that they were involved with the course proper. Remember, this improvement can be the product of experiences other than exposure to the text and participation in the course items.

The second procedure (QE 2), Test-Retest with Post Control, yields, in addition to the data of the types yielded by the simple Test-Retest, two sets of Post P values. The post test P values (Post P_c) are obtained from the responses of students who did not take a pretest. Therefore, they are free of any priming influence which may result from taking the course items before being exposed to the text. These data are also free of any practice effect score enhancement, so they should be representative of the data from typical student groups. A comparison of Post P_X will show the combined effects of practice and priming.

The third procedure (QE 3), Test-Retest with Dual Controls, in addition to the data of the types yielded by the second procedure, produces a set of P values from the first, or pre, administration, a set of P values from the second, or pseudo post, participation in the course items, and a set of differential P values. These will be designated Pre P_{CC}, Post P_{CC}, and Dif P_{CC}, respectively. Pre P_{CC} will be useful in evaluating the comparability of group X and group CC; Dif P_{CC} can be used to establish how much of Dif P_X is the result of both practice and incidental learning, without the text information. Post P_X - Post P_{CC}, will give an indication of the size of the performance increment that results from exposure to the text materials. Dif P_{CC} will indicate approximately how much of the score is due to practice effect and incidental, after priming, learning.

Table 2 summarizes the three procedures and the kinds of data available from them. There are also comments relating to the data.

We also realize that in some courses and situations some of these procedures are not practical. Many of the response forms, or answer sheets, used with correspondence courses are not readily adapted to automated response counting. Data from such forms can be hand counted or key punched for machine counting.

As the various P values can have various causes, there is no way that a computer can read the P values and accept or reject the items. A trained eye will always be needed to look at the several P values for each item and then at that item; afterwards, decisions can be made. An item with a high Pre P that has an instructional function should be retained. There also may be a reason for keeping an item with a very low Post P.

At this point, there are many unanswered questions. A paper such as this tries to open new avenues rather than supply pat answers.

Table 2

Data Acquired from Normal Procedure of Course Taking and from Three Experimental Designs

Procedure	Data	Comments
Normal	P	% students responding correctly
QE 1	Pre P	% students without training who respond correctly, related to non-course knowledge
	Post P	% trained students responding correctly
	Dif P	Increase in correct responses after training
QE 2	Pre P_x	Same as Pre P, QE 1
	Post P_x	Same as Post P, QE 1
	Dif P_x	Same as Dif P, QE 1
	Post P_c	Same as P, Normal procedure
	Post P_x - Post P_c	Increase in correct responses that result from pretest priming
QE 3	Pre P_x	Same as Pre P, QE 1
	Post P_x	Same as Post P, QE 1
	Dif P_x	Same as Dif P, QE 1
	Post P_c	Same as P, Normal procedure
	Pre P_{cc}	Same as Pre P, QE 1
	Post P_{cc}	% correct responses that result from training and priming that comes from taking the pretest without exposure to the text materials
	Post P_x - Post P_c	Same as in QE 2
	Pre P_x - Pre P_{cc}	Checks quality of groups X and CC
	Post P_x - Post P_{cc} Dif P_x - Dif P_{cc}	These show the effect that the text has upon making correct responses
	Post P_c - Post P_{cc}	Of little use in evaluating items

References

1. Campbell, Donald T., and Stanley, J. C., "Experimental and Quasi-Experimental Designs for Research on Teaching," Chapter 5 of Gage, N. L. (ed.), Handbook of Research on Teaching. Chicago: Rand McNally and Co., 1963.

2. Micheels, William J., and Karnes, M. R., Measuring Educational Achievement, Chapter 16. New York: McGraw-Hill Book Co., 1950.

THE EMERGENCE OF AN ITEM-WRITING TECHNOLOGY

Gale Roid and Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

Abstract

This paper provides a review of the emerging technology of test-item writing for criterion-referenced tests. Several different approaches to item development are discussed. A continuum of item-writing methods is proposed ranging from informal-subjective methods to computerized-objective methods. Examples of techniques include objective-based item writing, amplified objectives, item forms, facet design, domain-referenced concept testing and computerized techniques. Data from studies of item-writing techniques are also reviewed. Recommendations for further research and for applications to criterion-referenced testing are presented.

Revised version of a paper presented at the annual meeting of the Military Testing Association, Oklahoma City, October 1978.

THE EMERGENCE OF AN ITEM-WRITING TECHNOLOGY

Gale Roid and Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

Developers of any criterion-referenced (CR) achievement test have been confronted with the problem of writing test items which closely reflect the intent of instruction. This problem was earlier recognized by Osburn (1968) and by Hively and his colleagues (Hively, Patterson & Page, 1968). Bormuth (1970) was among the first to formally propose a science of item writing as a replacement for the informal, subjective experiences that commonly form the basis for item writing. In a review of Bormuth's approach to item writing, Cronbach (1970) remarked:

> The design and construction of achievement test items [illegible] been given almost no scholarly attention. The leading works of the generation--even the Lindquist Educational Measurement and the Bloom Taxonomy-- are distillations of experience more than scholarly analyses. (p. 509)

Since that time, there has been increasing activity in the area of item writing which clearly indicates the emergence of an item-writing technology that is grounded in theory and is now developing a research base.

The objective of this review is to describe the progress in the technology of item writing. This review should serve to stimulate theoretical and empirical work in the further advancement of item-writing technology, as well as provide useful guidelines to instructors and researchers who are interested in producing instructionally relevant achievement tests.

It is important, however, to provide an appropriate background for this review. Therefore, the steps one might employ in the construction of appropriate achievement tests are presented, and the role that tests play in systematic instruction is briefly described. Two distinct approaches to test-item writing are presented and contrasted. Studies are reviewed which bear on the feasibility and utility of each approach in producing effective CR tests. Finally, recommendations are offered for future research and development.

Developing CR Tests

Five steps are identified that are essential aspects of achievement test development (illustrated in Figure 1). These steps reflect a process which ideally occurs in any test development. The first step is the conceptualization of the content to be learned. Initially, the

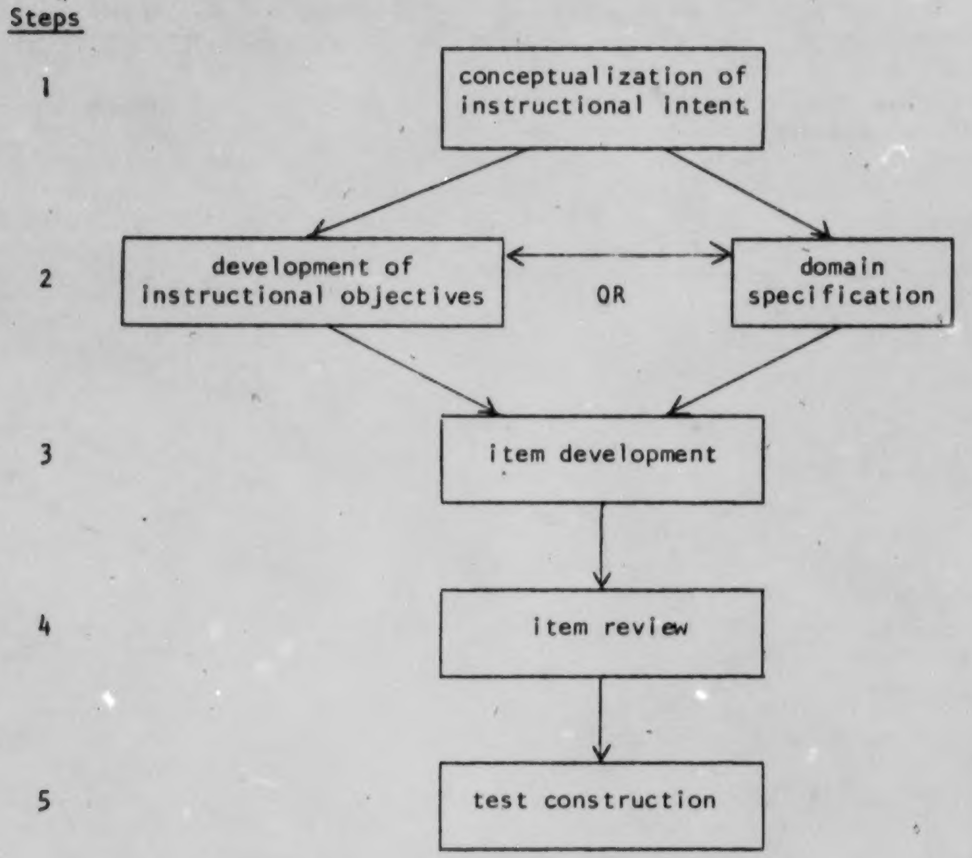

Steps
1
2
3
4
5
conceptualization of instructional intent
development of instructional objectives
OR
domain specification
item development
item review
test construction

Figure 1. Steps in the development of a CR test.

instructional developer or instructor must identify what the student must learn as a consequence of instruction. This step may be based on a task analysis or job analysis, or it may be admittedly introspective. It may be a "private event" that is purely abstract, but it is a vital beginning in the process of planning instruction and the associated CR testing that is part of this instruction.

The process of defining content has been the subject of much research. As Shavelson (Note 1) points out, there are at least three distinct ways in which to describe content structure. The first is hierarchically in the manner suggested by Gagné (1962), Ausubel (1963) or Bruner (1966). A second approach is content analysis whereby a system is used to categorize content. A third approach involves the defining of concepts and their relationships. However, the state of the science here appears to be more in the direction of informal, non-theoretical approaches to defining content, rather than the theoretical positions described by Shavelson.

Following the conceptualization of what is to be taught, in step two instructional intent is then transformed into either (a) instructional objectives which represent the behaviors to be elicited by learners as a result of instruction or (b) the specification of the content domain to be learned. Objectives, for quite some time, have been the mainstay of CR testing, although recent statements like those of Popham's (1975) and Millman's (1974a) have indicated that the inherent weakness in using objectives is that they permit considerable freedom when creating items, and studies like Roid and Haladyna (1978) offer empirical support to this view.

In the third step, items are developed using one of a number of item-writing techniques. The object in step three is to develop a universe or domain of test items which adequately represent the instructional intent as abstractly conceived in step one. A number of methods have been proposed and studied for developing items (e.g., the method of item forms developed by Hively, 1974), and the process is very much in line with the domain specification approach currently advocated by leading CR test theorists (Hambleton, Swaminathan, Algina & Coulson, 1978; Millman, 1974a; Popham, 1975).

While these item-writing procedures may generate items automatically, Haladyna and Roid (1978) and Hambleton, et al. (1978) have argued for processes whereby items are reviewed either by logical or by empirical means (step four). These item reviews are intended to identify defective items and either revise or discard such items before they are employed in CR testing. Thus, the resultant item domain is one in which logical and empirical reviews have been used to ensure the quality of the items.

The final step in test development (step five) is the selection of items for a CR test. While test blueprints and empirical item selection techniques have been advocated for years, there is strong evidence that random sampling of items should occur (Hambleton, et al., 1978; Popham, 1975; Haladyna & Roid, Note 2). Millman (1974b) provides some guidance on types of random sampling plans that may be employed to provide the adequate coverage of the content desired. The practice insures a high degree of content validity.

Within the area of CR testing there are many issues to be studied and resolved. These issues include (a) item review, (b) reliability, (c) decision making, (d) standard setting, and (e) validity. Each of these issues becomes the object of future study. However, the present review is focused on the first three steps of CR test development, which are related to item development. CR tests appear in a variety of educational settings, but the most appropriate of these settings would seem to be in instruction that is objective-based and systematic in nature.

Systematic Instruction and Systematic Testing

As test developers are aware, a good CR test is typically used in instruction to monitor student progress with respect to the intent of the instruction. There are a number of instructional systems, e.g., mastery learning (Bloom, 1968); personalized instruction (Keller, 1968; Robin, 1976) which treat instruction as an orderly process that is goal-based and student-centered.

As noted earlier, the CR tests that are developed are created by random sampling from a domain of items representing the instructional intent, as illustrated in Figure 1. An important distinction made by Millman (1974a) is that two types of CR tests exist--objective-based and domain-based. The objective-based CR test consists of items which were selected or written to reflect a single objective or a homogeneous set of objectives. The domain-based CR test is derived from a specification of the domain and rules for the development of items which do not permit a great deal of influence by the item writer on the item, thus greatly eliminating the potential for item-writer bias. Millman (1974a) has stated that the domain-based test is the purest form of CR test and a more desirable alternative to the objective-based test.

Within the framework of systematic instruction, there are several reasons for the increased attention to item-writing methods. Foremost among these is that most instructional systems need large collections of CR items in order to provide students with multiple forms for retests. When mastery is not achieved, instructors must provide suitable remediation and give retests until mastery is achieved. The consequence of this strategy, which seems to be common to virtually all forms of systematic instruction, is that a large collection of test items must effectively and logically represent instruction.

Another reason for increased attention to the development of items is the role achievement tests play in research. Educational researchers often must construct achievement tests to be used as dependent measures in their studies. Anderson (1972), in a classic paper, maintains that educational researchers tend to overlook the basic requirements of a system of measurement, "namely that there is a clear and concise definition of the things being counted" (page 145). This need can be extended to the area of evaluation research where the effectiveness of instructional programs is often determined by CR achievement tests that are not specifically described.

When item writers create items for CR testing using informal or subjectively inspired methods, they are likely to produce items which vary in quality and difficulty (Bormuth, 1970). The use of objectives or similar rules for item writing do not necessarily lead to better items. As demonstrated in a study by Roid and Haladyna (1978), the inherent subjectivity in item writing produces a bias that is difficult to overcome.

Another reason for concern with item development is that unless test-item writing methods are operationally-defined, these methods cannot be documented for other researchers or educators. If the test-item writer uses a mental process that cannot be described and communicated to another educator, the process of item writing remains a private event which is not defined and, hence, not replicable. An operationally defined method provides a precise description of how items are written so that two independent item writers using the same method produce virtually identical items. And these items have an integral link to instruction and a link to the intent of instruction.

Given this background, two fundamental approaches to CR test-item writing are identified and methods for writing or classifying items are described, and recommendations are offered for future research and development.

A CLASSIFICATION OF ITEM WRITING METHODS

All item-writing methods can be contrasted using a continuum which ranges from informal-subjective to computerized-objective (illustrated in Figure 2).

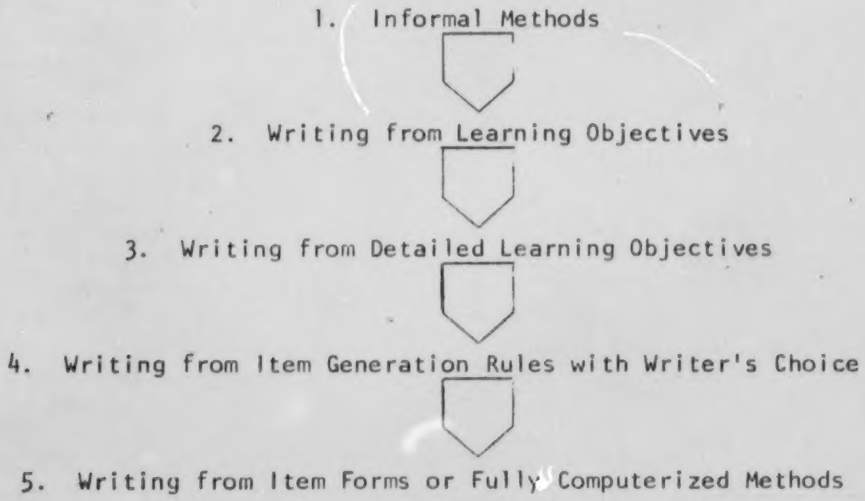

1. Informal Methods
2. Writing from Learning Objectives
3. Writing from Detailed Learning Objectives
4. Writing from Item Generation Rules with Writer's Choice
5. Writing from Item Forms or Fully Computerized Methods

Figure 2. A continuum of item-writing methods

The informal methods may involve a listing of the topics to be covered in the content of a course or may simply involve the instructor sitting down and writing items that are felt to be relevant to the course. At levels two and three of the continuum, learning objectives or detailed objectives may be written for a course of instruction, and these are used as guides in producing test items. At the fourth and fifth levels of the continuum, there may be a domain specification or universe of test items that is defined for a course of instruction and the tests that are used with it (Hively, 1974; Shoemaker, 1975). Since it is assumed that criterion-referenced tests are the appropriate tool for assessing student achievement in systematic instruction and because these tests are developed using either objectives or domains as the starting point, the emphasis in this review will be on these two major classifications of item-writing methods. The former subsumes levels two and three of the continuum, while the latter subsumes levels four and five.

Objective-Based Methods

Since the appearance of Mager's classic text, Preparing Instructional Objectives (Mager, 1962), there have been a plethora of basics dealing with the subject. The purpose in this section of the review will not be to show how to prepare objectives, but to evaluate the contribution of objectives to CR item writing.

Simply stated: "An objective is an intent communicated by a statement describing a proposed change in a learner--a statement of what the learner is to be like when he has successfully completed a learning experience" (Mager, 1962, p. 3). The key concept in this definition is the "intent" which is the raison d'etre for the objective. Given the objective, the test-item writer has a good idea what was intended, and is guided in developing CR test items which are appropriate to this intent. Further, objectives give organization to the content to be learned and are believed to provide focus to learning efforts. In fact, reviews by Duchastel and Merrill (1975), Hartley and Davies (1976) and by Melton (1978) indicate that the use of objectives does enhance learning, although the latter author warned that the perceived effectiveness of objectives is an oversimplification in light of the conditions that existed in the research on the effectiveness of objectives.

Studies dealing with item characteristics of CR tests were recently reviewed by Berk (Note 3) and by Haladyna and Roid (1978). These empirical studies, besides providing a technical base upon which item review may be performed, point to the deficiencies in the approach where objectives are used to generate items. In one study (Roid & Haladyna, 1978), two item writers used the same learning objectives as a guide in preparing items, but one item writer was found to consistently write more difficult items regardless of the objective. Thus, it seems that the very same subjectivity and bias that is present when the item writer uses his own intuitive notions can be present when objectives are used.

Dissatisfaction with the differences in items produced by item writers who use the objectives has prompted some to reject objective-based tests in favor of other "purer" forms of criterion-referenced tests. Popham (1978, p. 91) states: "The thrust of the emerging criterion-referenced measurement technology, therefore, is on increasing the capabilities of criterion-referenced tests to produce lucid descriptions of examinees performance." The objective-based CR test is viewed as a weaker form of a CR test in contrast to the domain-based CR test (Hambleton, et al., 1978; Millman, 1974a).

One solution to the problem of using objectives is the amplified objective, which is an elaboration of the objective, and which reduces uncertainty about the form and extent of items developed. An example is provided from Popham (1975, p. 147) which shows how an objective is transformed into an amplified objective (see Figure 3). The process thereby transforms an objective-based item-writing method into a domain-based method. That is, the amplified objective yields a pool of test items with well-defined characteristics.

Instructional Quality Inventory. Another approach to improving objectives is the Instructional Quality Inventory (IQI) developed by the Navy Personnel Research and Development Center, San Diego, and Courseware, Inc. (Ellis, Wulfeck II, Merrill, Richards, Schmidt & Wood, Note 4). IQI provides a method for examining the consistency between test items, objectives, and instruction. The IQI uses a matrix of test levels by content types that allows the test developer to classify test items and objectives in terms of both task and content. This examination of the relationship between objective and instruction is a major advance in the technology of item writing, although the degree to which the IQI has been successfully implemented is undetermined. Nevertheless, IQI provides a systematic approach to analyzing objectives which may allow for the creation of satisfactory items.

Classifying educational objectives. Two approaches to creating and classifying items and objectives will be briefly presented and reviewed. The first one is the most well known, the cognitive taxonomy proposed by Bloom and his colleagues (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956). A second approach is a typology introduced by Williams and Miller (1973).

Bloom's taxonomy consists of six categories ranging from knowledge, which deals with factual recall, to the highest level, evaluation, which involves judgment. The taxonomy has had tremendous impact on the thinking and practices of educators, and any discussion of objectives is incomplete without reference to this taxonomy. However, seldom are CR tests employed which involve this cognitive taxonomy. In a recent review of the properties of this taxonomy, Seddon (1978, p. 321) concludes, "No one has been able to demonstrate that these properties do

Descriptive Language: Concrete and Abstract Words Composition Skills

Objective: Given a sentence with a noun or verb omitted, the student will select from two alternatives the word that most specifically or concretely completes the sentence.

Sample Item:

Directions: Mark an "X" through one of the words in parentheses that makes the sentence describe a clearer picture.

Example: The racer (tumbled, went) down the hill.

Amplified Objective:

Stimulus Elements:

1. The student will be given simple sentences with the noun or verb omitted and will be asked to mark an "X" through the one word of a given pair of alternative words that more specifically or concretely completes the sentence.
2. Each test will omit nouns and verbs in approximately equal numbers.
3. Vocabulary will be familiar to a third- or fourth-grade pupil.

Response Alternatives:

1. The student will be given pairs of nouns or pairs of verbs with distinctly varied degrees of descriptive power.
2. In pairs of verbs, one verb will either be a linking verb or an active verb descriptive of general action (e.g., is, goes), and one verb will be an action verb descriptive of the manner of movement involved (e.g., scrambled, skipped).
3. In pairs of nouns, one noun will be abstract or vague (e.g., man, thing), and one noun will be concrete (e.g., carpenter, computer).

Criterion of Correctness:

The correct answer will be "X" marked through the more concrete, specific noun or through the more descriptive action verb in each pair.

Figure 3: Example of an amplified objective.

not exist. Conversely, no one has been able to demonstrate that they do." Thus, the utility of Bloom's taxonomy as a tool for CR test developers is still questionable.

A typology for test questions was originally introduced by Williams and Miller (1973) which viewed objectives and items particularly as representations of one of five possible types. The term "typology" was used as no order for these categories was implied. A fuller treatment of this work (Miller, Williams & Haladyna, 1978) reveals a system much like Bloom's which, instead, focuses on verbs in test questions as keys to interpreting the cognitive category of behavior in which a test item falls. The categories include factual recall, summarizing, predicting, evaluating, and applying. The example in Figure 4 provides a brief definition of each level and examples of questions at each level. Williams (1977) added another category to this typology, instantiation, which is a derivative of summarizing. His empirical study of the typology revealed that students with a minimum of training could classify test items with a high degree of accuracy. Like other systems for classifying objectives and items, there is no empirical research to support its use, other than Williams' study, and its usefulness at various educational levels and content areas is unknown.

ITEM DEVELOPMENT BASED ON DOMAIN SPECIFICATION

The concept of "domain-referenced" testing was first reported in 1968 (Hively, Patterson & Page, 1968; Osburn, 1968) and further developed by Hively and his colleagues (Hively, Maxwell, Rabehl, Sension & Lundin, 1973). As we reported earlier, domain-based CR tests are derived from content specifications as opposed to objective-based tests which are derived from instructional objectives.

A new technology of domain specification provides an alternative to objective-based item-writing methods. There are at least five distinctly different approaches to item creation which involve domain specifications. These include: (a) item forms, (b) linguistic-based approaches, (c) facet theory, (d) concept-based testing, and (e) computer-based methods. Each is described, and research is reviewed which bears on the success with which the method has been employed in CR testing.

Item Forms

Items for domain-based tests may be written from specifications that describe the format and even some of the wording of the resulting items. The specifications are called "item forms" (Hively, 1974), and the pool of items that an item form creates is the domain to be assessed.

"An item form," explains Osburn (1968, p. 97), "has the following characteristics: (1) it generates items with a fixed syntactical

Cognitive Type	Definition	Syntactical Forms	Example of a Question
Factual Recall	The reproduction of a stimulus element exactly as it was presented.	Name State Describe	When did Columbus discover America? a. 1492 b. 1489 c. 1776
Summarizing	The understanding of concepts and the tendency to correctly identify examples, instances, or attributes of the concept.	Identify Define Translate Typify Represent Describe	What is a good example of alliteration? a. gurgling b. school - pool c. blue - blood d. up - down
Predicting	The use of rules in contingent relationships. The student is given a situation and must anticipate a consequence which is based on a rule.	If..., then....	If the temperature of the fluid in the flask exceeds 100° C, then a. all fluids will evaporate. b. the mixture will explode. c. nothing will happen.
Evaluating	The tendency to (a) select a criterion or criteria, (b) to use a criterion, or (c) both select and use a criterion for a decision.	Which is best, worst; highest, lowest; most, least...?	From the standpoint of efficiency, which procedure is best? a. driggling b. harpoling c. craterling d. quarboling
Applying	Problem solving which involves the "how to" of applying involves (a) sensing the problem, (b) defining the problem, (c) selecting principles, rules, or methods by which the problem is solved, and (d) selecting or generating solutions.	No standard forms.	To achieve a well-balanced city water system, which plan will provide a steady supply of water in all seasons? a. a deep well system west of town b. a deep well system east of town c. a reservoir in west hills d. a pipeline from neighboring Independence

Figure 4. Definition, syntactical structure of questions, and examples of a cognitive typology of CR test items.

structure; (2) it contains one or more variable elements; and (3) it defines a class of item sentences by specifying the replacement sets for the variable elements." The item forms developed by Hively and Osburn were in science or mathematics. For example, the following is an item form for a basic mathematics concept:

Item Wording: Which of the following numbers is a prime number?

(a), (b), (c), (d)

Elements to Complete the Item: Four numbers, (a) to (d) are provided and the student is required to check the one that is the prime number. These numbers must be two or three-digit integers, and all must be odd numbers. The foils must be non-trivial as defined by the fact that they should be factorable into a minimum of 3 factors.

Sample Item: Which of the following numbers is a prime number?

27, 31, 147, 189

Correct Answer: 31

No studies have been observed that deal with the feasibility of item forms or empirical comparisons between this approach and others. Hively, Patterson and Page (1968) studied the empirical properties of items developed from item forms using Cronbach's generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Results of the study were promising in that items were produced that showed response patterns that suggested distinct and homogeneous classes of behavior.

The most significant work to date on item forms was the five-year cooperative project, MINNEMAST (Hively, et al., 1973). The monograph documents a domain-based test development from item forms, providing a rich resource of examples and problems encountered. Foremost among these problems is the extraordinary cost in the development of item forms and the administration and scoring of test items, many of which were not machine scorable.

Further, there are concerns expressed for the feasibility of such an approach (Popham, 1975, p. 136). Until item forms can be made more efficient, their potential may be limited to subject matter that is more objectively structured and identifiable.

Millman and Outlaw (1977) have recently implemented item forms in the tests used for several college courses. They have developed a special programming language for a small-computer system that allows an item writer to construct an "item program." This is a computer program that directs the system to produce multiple questions. The item program defines a structure for each question. Most of the wording of the item can be fixed and parts of the item can be variables that are replaced to create unique questions. Variable elements can be words or random numbers or quantities that are mathematically computed. An example of a very simple item form and an item program for a math problem is shown in Figure 5. One advantage of this system is that only the item program needs to be stored, not the individual items. A test can be constructed and printed by the computer.

Item Form

"How much is (X) plus (Y)?"

Where X and Y are integers from 1 to 10.

Item Program

```
10 LET X = RANDOM (1, 10)

20 LET Y = RANDOM (1, 10)

30 QUESTION CONTENT " How much
     is ", X, " plus ", Y, " ? "

40 ANSWER CONTENT  X + Y
```

Figure 5. Item form and item program from Millman and Outlaw (1977)

Linguistic-Based Approaches

Bormuth (1970) was the first to describe a technology of item writing for assessing learning from prose material. He described rules that are a series of directions which tell an item writer how to transform segments of prose instruction into questions. Bormuth outlined two types of transformations: (a) items derived from sentences and (b) items derived from the relationships between sentences (1970, pp. 39-55). An example of sentence-derived items that assess the recall of prose material are those created by the "wh-transformation." These items would be written using a detailed set of rules summarized as follows: "Select sentences from the instructional materials, replacing a 'wh' word such as <u>who</u>, <u>what</u>, or <u>where</u> for the appropriate part, e.g., subject noun, in each sentence." For instance, "The test developer computes the validity coefficient," could be transformed to: "Who computes the validity coefficient?" These are particularly useful because they can be written to assess learning of each of several ideas in one sentence, and can be made into either completion or multiple-choice format.

Through the use of paraphrasing, sentence-derived items can also be written to test comprehension of prose material. Anderson (1972) has emphasized the importance of paraphrasing and has defined it as the case where (a) all substantive words in a sentence are replaced and (b) the original and paraphrased sentences have equivalent meaning.

Questions can be developed from the relations between sentences; for example, by questioning the cause of an action described in a prose passage. For instance, the sentences (a) Jim hurt his hand, (b) He was cleaning his knife, and (c) His knife accidentally slipped, can be examined for implied causation, resulting in the question, "What caused Jim's hurt hand?" (Bormuth, 1970, p. 54).

Finn (1975; Note 5) and Roid and Haladyna (1978; Note 6) have extended the work of Bormuth by developing multiple-choice item-writing methods for prose learning. Finn's original work (1975) involved a rather lengthy, 82-step algorithm. A streamlined version of this algorithm was developed by Roid and Finn (Note 7) and included the following important steps:

1. Analyzing the text;
2. Selecting sentences by keywords;
3. Transformation of sentences into questions; and
4. Generation of foils for multiple-choice format.

Analyzing the text. To develop questions that measure important aspects of a prose passage requires a selective screening of the text. One approach to screening the prose material is to use a team of teachers or curriculum experts to identify the "instructionally relevant" sentences. Many instructional programs include sentences that are directions to the student, references to illustrations, or other verbal information that is not directly related to the learning objectives for the program. Screening of this material by consensus would be essential for the creation of meaningful, relevant items.

Another approach to text analysis was proposed by Finn (Note 5), who used a word-frequency analysis of a prose passage. A prose passage is screened by (a) counting the number of times that each noun or adjective appears in the passage and (b) identifying the standard frequency index of each noun or adjective. The standard frequency index of each word is a numerical estimate of how often the word occurs in a large sample of words from American textbooks (Carroll, Davies & Richman, 1971). The Carroll, Davies and Richman book or its computer-tape version can be used to get the standard frequency index of each word in the passage. The word "the" has the highest standard frequency index of any word, because the average American student is likely to encounter the word "the" once in every ten words in a textbook. The word "incarnation," for example, has the lowest index because the average student is likely to encounter this word less often than once in every billion words.

Selecting sentences by keywords. One approach to identifying the important sentences in a passage would be to have instructors or content experts underline the key sentences in the passage. A consensus of markings can be used to identify the most important sentences. If this consensus is based on learning objectives for the prose material, the method becomes objective-based. The sentence-transformation methods to be described are, therefore, domain-based for two reasons: (a) sentences can be randomly sampled and (b) transformations can be operationally defined.

The approach to identifying keywords in prose proposed by Finn (Note 5) is to identify "high information" words--words that are relatively rare in American English and occur infrequently in the passage. The sentences in which these high information words occur can then be sampled for transformation into items which assess important information in the passage. Degree of information in this context is measured by the amount of uncertainty in the meaning of a sentence that is created if a word is deleted. High information words are those that are difficult for students to guess if they are deleted from sentences such as is done in a Cloze test (Culhane, 1970). Cloze tests are completion tests in which every fifth word has been deleted from a prose passage. The task for the student, then, is to fill in the missing words. The easiness with which a word is guessed by a student is a measure of the amount of information it provides. The task in a Cloze test is similar to the exemplary problem in information theory (Shannon & Weaver, 1949) where a person is receiving a message, but because of noise on the channel, he is not always sure which message he hears (Finn, Note 8). The information in a garbled message is a function of the amount of doubt the receiver has about its meaning and is related to the probability of occurrence of certain words or letters. A missing word which is a common word in the English language would give less information, because students would more easily guess that it completes a sentence.

Finn (Note 8) has shown that the easiness with which a word is guessed on a Cloze test is predicted by two important measures derived from a word-frequency analysis of a passage: (1) the standard frequency index and (2) text frequency. Words that have a low standard frequency index (infrequent in American textbooks) are defined as high in information. However, there is one case in which the information of these words is reduced in relation to a given passage. If the word is repeated frequently (i.e., it has a high text frequency), the information value of that word is reduced and students will guess it more often in a Cloze test following reading of the passage. In other words, repetition of a word, even if it is rare in American English, lowers its information value. Candidates for good question words are those which are both rare in American English (have a low standard frequency index) and occur infrequently in a prose passage.

Not all parts of speech are equally good question words, even though they may be high information words. Verbs and adverbs, in particular, require difficult transformations when removed from a sentence. For example, the sentence, "Finn argued the point made by Bormuth," when transformed to "What did Finn do to the point made by Bormuth?" seems clumsy and seems to be a less important question than the question "Who argued the point made by Bormuth?" According to Roid and Finn (Note 7), the most promising parts of speech are adjectives and nouns, or phrases that contain them.

Transformation of sentences into questions. Once an important word and sentence have been identified for a question, the sentence must be examined and prepared for transformation. Some sentences include references to previous sentences, e.g., "This implies that . . . ". A phrase from the previous sentence must be inserted into the place of the referent (e.g., in place of "This"). Also, sentences that are compound or that contain long clauses that introduce more than one idea into the sentence need to be separated. The portion of the sentence containing the question word is separated and used by itself if possible (see Finn, Note 5, for guidelines).

The next step is to eliminate the question word and to transform the sentence into a question. The question word, usually an adjective, a noun, or its phrase, is removed and is replaced with a wh-word. Where several wordings are possible, an attempt is made to stay as close to the wording of the original sentence as possible, unless paraphrasing is being used.

Sentence transformations do not produce 100% agreement among item writers in all cases because of such things as the replacement of phrases from previous sentences. Finn (1975, pp. 357-363) discusses some of the discrepancies among item writers. One study (Roid, Haladyna, Shaughnessy & Finn, Note 9) showed that differences between item writers were not statistically significant when the Finn method was used.

Generation of foils. As is common knowledge among item writers, the writing of good foils for multiple-choice questions is challenging. The first step in an algorithmic generation of foils is to classify the question words so that possible foils can be obtained from a list of words in the same classification. The most logical source of foil words would seem to be from the prose passage itself.

Roid and his colleagues (Roid & Finn, Note 7; Roid, Haladyna, Shaughnessy & Finn, Note 9) developed a technique for algorithmic foil construction. The algorithm uses words from the prose passage itself as foils. Two variations of the algorithm were developed: one for nouns and one for adjectives.

In the case of nouns, those with a standard frequency index of 60 or less were semantically classified using the method of Frederiksen (1975). For example, some nouns were classified as concrete inanimate nouns. For a given question word, a random sample of three other nouns from the passage that were similarly classified were drawn to create foils.

In the case of adjectives, research on semantic differential technique was used as a basis for classifying adjectives from the passage (Nunnally, 1967, pp. 536-638). In semantic differential research, three factors are usually identified: (1) *evaluation*, such as "good," "bad," etc., (2) *potency*, such as "strong," "weak," etc., and (3) *activity*, such as "quick," "slow," etc. In addition, Nunnally has defined a fourth factor, "*familiarity*," such as "simple" or "complex." These four factors were used to classify the adjectives in the passage that had standard frequency indexes of less than 60. As a further screen of the adjectives, the Dale-Chall List of 3,000 Familiar Words (Dale & Chall, 1948) was used. The adjective needed to be absent from that list so that extremely common adjectives could be eliminated. These common adjectives were suspected to be too easy as foils.

Research on these linguistic-based methods for generating items has been mainly limited to a series of studies which trace the development of a method for creating appropriate multiple-choice CR test items. The first of these studies, as described earlier, contrasted objective-based and informal methods of item writing (Roid & Haladyna, 1978). The objective-based method was more in keeping with the amplified objectives approach. The results of this study showed that while objectives provided guidelines in the preparation of items, one item writer's items were about 10% more difficult than the other's. The consequence of developing CR test forms based on any particular item writer can be great with respect to misclassifying student examinee performance as adequate or inadequate, as is typically done in many forms of systematic instruction. Thus, item-writer bias is a phenomenon that affects the difficulty, if not the quality, of CR test items produced informally or with objectives. This study suggests that CR test-item writers should not proceed from step one, conceptualization, to step 3, item writing, or use loosely-stated objectives, as suggested in step 2.

In the second of this series of studies, Roid and Haladyna (Note 6) examined the effects of variations in linguistic-based algorithms on item characteristics including instructional sensitivity, a criterion measure based on the tendency for items to exhibit change in difficulty as a function of instruction.[1] Four item writers were compared on two methods of selecting sentences, two types of question words, and two foil construction methods. No significant differences between item writers were found on item difficulties, indicating an absence of item-writer bias. Keyword nouns, which are relatively rare words in American textbooks that appear frequently in a prose passage, were found to be unacceptable as question words. Algorithmic methods of foil writing were found to be feasible. Thus, this study indicated that item-writer bias could be eliminated through the use of certain rules dealing with the way sentences are identified and transformed into multiple-choice questions. The study also points to a need for further work that is needed in the refinement of the algorithms in an effort to achieve more fully automated or even computerized procedures.

---

[1] See Haladyna (1974) and Haladyna and Roid (1978) for fuller discussions of this characteristic and measures of it.

In a more recent study in this line of research, Roid, Haladyna, Shaughnessy and Finn (Note 9) examined some of the refinements developed as a result of the previous study in contrast to methods of item writing that are based on paraphrasing of keywords. It was found that passages with greater density (i.e., sections that provided more information) yielded harder items. More importantly, letting item writers have more freedom in the selection of foils produced better items based on the criterion of instructional sensitivity. Verbatim transformations led to items with higher instructional sensitivity. The procedures examined also led to greater control of item difficulty than previously observed. This study documents some of the intriguing advantages of algorithmic multiple-choice item-writing methods, and this study also points to the need to further study and improve item-generating techniques based on Bormuth's theory of achievement testing. The goal of reducing item-writer bias was effectively achieved, and future work will concentrate on making the process more cost effective and efficient.

Facet Theory[2]

Structural facet theory (Foa, 1958) has existed for some time and has mainly served as a research tool, particularly in the area of attitude measurement. Only recently have there been applications to CR test construction (Engel & Martuza, Note 10; Berk, Note 3). The purpose of facet theory is to provide the structure and boundaries of a domain of testing conditions. For this reason, facet theory is viewed as a method for developing a population of items representing the domain of instruction. The primary advantage of facet theory is that the analysis of content has semantic meaning in a theoretical sense, and there is no need to conduct empirical analyses to search for meaning. The logical analysis of sentences leads to meaningful test items which are easily interpretable. That is not to say that empirical observation is unnecessary in facet theory, but that the theory a priori specifies the nature of the material to be learned and tested. Thus, facet theory, like other similar approaches, allows for an objective specification of the domain which is the target of instruction.

Facet theory specifies the limits of the domain and the orderings of its subparts. In the theory, two aspects are hypothesized: content and statistical. With content, the domain is specified using a semantic structure called the "mapping sentence." The content structure is the framework for predicting the statistical structure, which is later tested using observations, thus enabling the user to relate theory to observation. In the context of achievement testing, the mapping sentence is a mechanism for defining a content domain and a related set of test items to measure achievement in that domain.

---

[2] The source of the information presented here was mainly derived from excellent presentations of facet theory by Engel and Martuza (Note 10) and by Berk (Note 3). The reader is referred to these original sources for a more detailed treatment of facet theory for CR tests.

The mapping sentence has fixed and variable parts resembling an item form shell. Parts of the sentence called "facets" are identified which represent some specific information for testing, and the facet elements operate in much the same manner that replacement sets operate in item forms. The sum of all desirable patterns of facets constitutes a facet design.

Using an example provided by Berk (Note 3, p. 2):

A. Domain Specification Strategy
   1. Sentence transformation
   2. Item forms
   3. Algorithms
   4. Mapping sentences

are most appropriate for defining

B. Content Domain
   1. Reading
   2. Language
   3. Mathematics
   4. Science
   5. Social studies

these content domains

The mapping sentence for this example would have two facets. The elements of each facet are ordered in some meaningful way. There is a set of rules for choosing facets and their elements (McGrath, 1967). These rules are summarized from Berk's paper (Note 3):

1. Objects should be classified by all properties or facets.

2. Each facet should be divided into an exhaustive set of categories or elements.

3. The elements should be mutually exclusive; that is, each element is classifiable into one and only one category.

4. The logical relationship among elements of a facet should be specifiable.

5. The logical relationship among facets should be specified.

6. The facets should exhaust the domain of interest.

From the example given above, there are a total of twenty combinations of elements from Facet A and Facet B. Thus, twenty statements exhaust the domain of possibilities, and twenty true-false items are possible. For example: Sentence transformation is most appropriate for defining mathematics. This is A1B3, which is a false statement.

This particular facet design is useful for true-false or completion formats. Building a multiple-choice domain of items is considerably more complex. Besides identifying the correct answer for a particular facet, the additional burden is to produce three or four plausible foils. This process, described by Berk (Note 3), involves a logical analysis of potential distractors which are drawn from the elements of the facet. Using the previous example with some modification:

Item transformations are most appropriate for defining:

a. social studies
b. language
c. reading
d. mathematics

The benefits of facet designs, as well as other similar approaches, were discussed by Engel and Martuza (Note 10):

1. Both item stem and foils can be systematically constructed.

2. Facet design is based on a theory of content and how content is defined. Therefore, the identification of foils occurs in the context of how foils are more or less attractive as incorrect responses. As a consequence, incorrect responses have meaningful interpretations in a diagnostic vein.

3. The procedures provide a logical connection between content and the multiple-choice item.

4. Items produced may be logically compared with respect to content difficulty and appropriateness, thus making the construction of parallel test forms easier and less subject to capriciousness which exists when random sampling is used to create items.

As noted earlier, facet theory is a relatively young field of content specification for domain-based CR tests. There is very little known about its applicability to various subject matters or the empirical characteristics of tests constructed using facet designs.

Engel and Martuza (Note 10) conducted an empirical study of facet designs. The procedures led to functionally equivalent parallel forms tests. Further, results indicated that the method works equally well with highly structured material like mathematics as well as more abstract material. Finally, the study showed the feasibility of facet theory as a method for domain-based CR test-item construction. It is also interesting to note that like amplified objectives, the facet design can be used with an objective as the mapping sentence, thereby capitalizing on existing objectives.

Like other approaches to domain-based testing, facet theory appears quite promising. These seminal works by Engel and Martuza (Note 10) and Berk (Note 3) provide a clear picture of the nature and potential of facet theory. There remains, however, much work to be done to refine the theory and to apply it to various subject matters as well as to compare it to other domain-based approaches in an effort to uncover which set of procedures is most efficient, feasible, and defensible in light of the abstract conceptualization of instructional intent posited as the first step in CR test development. As with other approaches, more development coupled with empirical research should reveal the utility of facet theory and its eventual role in CR testing.

Concept-Based Testing

The work of Markle and Tiemann on the teaching and testing of concepts can be used to create domain-based tests that go beyond the factual level of learning (Markle, 1975; Markle & Tiemann, 1974; Tiemann & Markle, 1978; Tiemann, Kroeker & Markle, Note 11; Tiemann & Markle, Note 12). They have defined concepts as classes of objects, events, or relations which vary among themselves and yet are all grouped together and called by the same name. A student's understanding of a concept is tested by checking for generalization to new examples and discrimination of nonexamples. A set of examples and nonexamples that are different from those used in teaching are used to test the student's understanding of the concept. If we were teaching the concept "chair," we might use the examples of a metal kitchen chair and an upholstered chair and the nonexamples of a stool and a church pew in the teaching exercise. In testing for understanding of the concept of chair we might use the examples of a rocking chair and a rattan chair, and the nonexamples of a bench and a love-seat.

Tiemann and Markle (1978) provide guidelines and many practical examples of the analysis of concepts. The analysis of concepts involves listing the variable and critical attributes of the concept. A variable attribute is a property of any particular example which can be varied without changing an example to a nonexample. For instance, the number of legs is variable in the concept chair, because we can have a modernistic chair with a pedestal or a standard four-legged chair. Critical attributes are true for every example of the concept, and if they are removed, the example becomes a nonexample. For instance, the requirements of a "single-person seat," a back and a rigid seat, are the critical attributes of chair. Variable attributes are whether it has rockers, arms, the material it is made of, etc. After the critical attributes and variable attributes have been listed, and lists of examples and nonexamples have been written, it then becomes possible to construct domain-based criterion-referenced tests for a given concept. Such a test would be constructed by choosing a random sample of examples and nonexamples and systematically varying critical attributes and variable attributes. An example of a concept analysis and a sample item for the concept "antonym" is given in Figure 5 from Tiemann and Markle (1978).

Markle and Tiemann recently extended their work to multiple coordinate concepts (Tiemann, Kroeker & Markle, Note 11). An example of coordinate concepts is the four behavioral concepts of positive and negative reinforcement and positive and negative punishment. In this case, the four concepts interrelate to the point where an example of one concept is a nonexample of the other concept. Students need to learn to reject a nonexample in one case but accept it as an example in the other case. The Tiemann, et al., study (Note 11) provides an example of how a domain-based test is produced for a set of four coordinate concepts by systematically sampling examples of each of the concepts and varying them on their attribute dimensions.

Like other emerging technologies of test-item writing and domain specification, there is little empirical work to support the approach being advocated. The concept-based approach of Markle and Tiemann provides a level of cognitive questioning that goes beyond levels typically assessed by the linguistic-based approach. And concept-based testing also serves to capture areas of instructional intent that are not handled very well by item forms which seem most applicable to discrete objects such as those found in mathematics, science, and basic skills (e.g., spelling).

Computer-Based Methods

Computers have been used for many years as aids in assembling or administering tests (e.g., Atkinson & Wilson, 1969). Early attempts centered on the use of item banks containing all of the actual items from which samples were drawn for testing. More sophisticated systems included the composition of items such as was done in the mid-1960's in the drill-and-practice exercises of the Stanford computer-assisted instruction project (Suppes, Jerman & Groen, 1966). Computer programs with the capability of generating items can be used to create domain-based tests, and, for that reason, they will be described in more detail.

The major author languages used in computer-assisted instruction have the capability of producing algorithms for domain-based test items. Several of the CAI languages discussed by Roid (1974) such as COURSEWRITER, PLANIT and TUTOR have functions which allow an item form to be programmed as described by Millman and Outlaw (1977). For example,

Grammar Concept: Antonym*

A word which:

Critical Attributes

1. has a meaning opposite to the meaning of some other (given) word

2. is the same part of speech as the given word

3. is a new word, not a variation of the given word

Variable Attributes

4. may be drawn from various parts of speech:
   a) nouns
   b) verbs
   c) pronouns
   d) adverbs
   e) adjectives
   f) prepositions

5. relative syllabic length of two words may be:
   a) equal
   b) unequal

6. opposition of meaning may exist:
   a) across some continuum
   b) in a dichotomous sense

Teaching Examples

1.	bad; good	4e,5a,6a
2.	danger; safety	4a,5a,6a
3.	live; die	4b,5a,6b
4.	he; she	4c,5a,6b
5.	rapidly; slowly	4d,5b,6a
6.	in; out	4f,5a,6b

Teaching Nonexamples

1.	vain; greedy	lacks only 1
2.	reason; motive	lacks only 1
3.	we; us	lacks only 1
4.	above; upon	lacks only 1
5.	merrily; sad	lacks only 2
6.	happy; unhappy	lacks only 3
7.	capable; incapable	lacks only 3
8.	disputable; agree	lacks only 2

Testing Examples

1.	hot; cold	4e,5a,6a
2.	loss; gain	4a,5a,6a
3.	elevate; lower	4b,5b,6a
4.	you; me	4c,5a,6b
5.	gaily; sadly	4d,5a,6a
6.	over; under	4f,5a,6b

Testing Nonexamples

1.	imaginary; fanciful	lacks only 1
2.	chair; couch	lacks only 1
3.	behind; next to	lacks only 1
4.	gloom; bright	lacks only 2
5.	violent; non-violent	lacks only 3
6.	valid; invalid	lacks only 3
7.	weak; forcibly	lacks only 2

Sample Test Item

Which of the following pairs of words are antonyms?

a. imaginary -- fanciful
*b. elevate -- lower
c. valid -- invalid
d. weak -- forcibly

Correct Answer: b

Figure 5. Example of a concept analysis used to develop domain-referenced tests of concept learning.

*Adapted from P. W. Tiemann and Susan M. Markle. Analyzing Instructional Content: A Guide to Instruction and Evaluation. Champaign, IL.: Stipes Publ., 1978, p. 257. By permission of the publisher and authors.

Atkinson (Atkinson & Wilson, 1969, p. 153) used COURSEWRITER to create reading exercises and criterion tests for the Stanford reading programs. A sample exercise is the sentence, "Jan saw the _____ hat," for which the student is to choose one of a set of computer-assembled words, such as "tan," "fat," "man" or "run." The fill-in answers are selected by rules from words previously presented in lessons.

Another example is the work of Fremer and Anastasio (1969) who used computers to help generate items for testing spelling. They conducted an analysis of types of misspellings used by writers of spelling items. A set of error generation rules were developed and programmed for a computer. Error generation rules included the inversion of letters within a word, omission of letters, or insertion of letters. An example for the word "preferable" would be "perferable" or "preforable" or "preferabal." Fremer and Anastasio found that computer-generated lists of spelling items were judged highly useful by a panel of spelling test developers.

Beginning with the pioneering work of Hively, Patterson and Page (1968) and Osburn (1968), a great deal of work has been done on domain-based tests in mathematics. For example, Hsu and Carlson (1973) developed routines used to construct tests for the elementary mathematics level of the Individually Prescribed Instruction program. They used the concept of item forms developed by Hively in programming item-generation routines. Hsu and Carlson make the important suggestion that statistics for item forms be computed by collecting data from tryouts of each item form. Because individual test items are automatically produced, the best way to insure the quality of test items is to improve the quality of the item forms. By field testing and keeping statistics at the item form level, it will be possible to develop higher quality domain-based tests.

Beginning with the reported work of Osburn (1968), a number of university professors have developed computer-generated testing systems, particularly in the sciences. For example, Johnson (1973) has developed a system for computer-generated items for chemistry at the college level. Each of a series of subroutines defines an item form. These item forms include numerical constants which are randomly generated by computer or variable wordings which might include different names of chemical compounds. An example of an item form and an individual item from this system is given in Figure 5.

Military uses of computerized item writing include the work by Braby, Parrish, Guitard and Aagard (Note 13) at the Orlando Naval Training Center. They developed algorithms for teaching and testing symbol-recognition.

Other computerized item-writing efforts have been described by Olympia (1975) and Vickers (1973). The work of Vickers is interesting in that it involves the computer generation of items useful in the teaching of FORTRAN programming. A series of subroutines, that employ

Item Form

IF _____ ML. OF _____ MOLAR _____ IS MIXED WITH _____ ML. OF _____ MOLAR _____, THE FINAL SOLUTION WILL BE:

A) _____ MOLAR IN _____

B) _____ MOLAR IN _____

C) _____ MOLAR IN _____

D) _____ MOLAR IN _____

E) _____ MOLAR IN _____

Sample Item

IF 43.6 MLS. OF 1.50 MOLAR NAOH IS MIXED WITH 38.5 MLS. OF 1.14 MOLAR HNO3, THE FINAL SOLUTION WILL BE:

A) 1.33 MOLAR IN OH (-)

B) 1.10 MOLAR IN H (+)

C) 0.260 MOLAR IN H (+)

D) 1.33 MOLAR IN H (+)

E) 0.260 MOLAR IN OH (-)

Figure 5. Example of an item form and item from Johnson's Computer-Generated Repeatable Chemistry Exam System (Johnson, 1973).

random numbers, are used to compose FORTRAN-like statements. Then, the student is asked to discriminate between correct and incorrect statements or to classify types of variables. This is an excellent example of a sophisticated item-writing method that produces a large domain of items.

In summary, there is a wide variety of applications of computerized item-writing methods. Many of these methods are in use in military, college, and university courses, particularly in the sciences. The capability to implement these methods is available at most of the major computer centers in the nation. Thus, the technology is available for writing domain-based criterion-referenced tests. The challenge that remains is in the specification of domains and the definition of item-writing algorithms in a wide variety of subject-matter areas. Also, more creative efforts are required to develop domains at the conceptual and higher levels of learning following the recommendations of Tiemann and Markle.

CONCLUSIONS AND RECOMMENDATIONS

Instructional development has been capably served by several principles of learning and testing which involve the use of instructional objectives and other testing aids. Research on variables within systematic instruction (reviewed by Block, 1971; Duchastel & Merrill, 1973; Hartley & Davies, 1976; Melton, 1978; Robin, 1976) has been impressive. Both systematic instruction and the use of instructional objectives appear to have positive effects on learning. Unfortunately, objectives permit much too much freedom to the often inexpert item writer which in turn results in many items which are instructionally irrelevant or psychometrically unsound. Despite logical and empirical methods of item review, many of the problems in producing items may be avoided by employing one of several domain-based item-generating approaches. The five reviewed here are sound in theory, research and development. Preliminary findings indicate a vast potential for the creation of large groups of items which may form the basis of sound CR testing in the future.

The key to all this activity is the acceptance of the process in CR test development illustrated in Figure 1. Test theorists and practitioners are in accord when they maintain a concern for a logical and close correspondence between instruction and testing. Because domain-based methods elaborate on objective-based methods, it is possible to achieve almost perfect objectivity in the creation of test items. Therefore, the item-writing methods being reviewed promise a more scientific approach to item development, which in turn improves the measuring of student achievement. This improvement should, in turn, help instruction to more closely and accurately monitor student progress while educational researchers should find the work of creating achievement tests as research tools more fruitful.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Atkinson, R. C., & Wilson, H. A. Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

Ausubel, D. P. Cognitive structure and the facilitation of meaningful verbal learning. Journal of Teacher Education, 1963, 14, 217-221.

Block, J. H. (Ed.). Mastery learning: Theory and practice. New York: Holt, Rinehart, and Winston, Inc., 1971.

Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1, 1-12.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of educational objectives. New York: Longmans, Green, & Company, Inc., 1956.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bruner, J. S. Toward a theory of instruction. Cambridge: Harvard University Press, 1966.

Carroll, J. B., Davies, P., & Richman, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

Cronbach, L. J. Review of "On the theory of achievement test items." Psychometrika, 1970, 35, 509-511 (Book Review).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley, 1972.

Culhane, J. W. CLOZE procedures and comprehension. The Reading Teacher, 1970, 23, 410-413.

Dale, E., & Chall, J. S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-28.

Duchastel, P. C., & Merrill, P. F. The effect of behavioral objectives on learning: A review of empirical studies. Review of Educational Research, 1973, 43, 53-69.

Finn, P. J. A question writing algorithm. *Journal of Reading Behavior*, 1975, *4*, 341-367.

Foa, U. G. New developments in facet design and analysis. *Psychological Review*, 1965, *72*, 272-274.

Frederiksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology*, 1975, *7*, 371-458.

Fremer, J., & Anastasio, E. J. Computer-assisted item writing -- I (Spelling items). *Journal of Educational Measurement*, 1969, *6*, 69-74.

Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, *69*, 355-365.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 93-99.

Haladyna, T., & Roid, G. *The role of instructional sensitivity in the empirical review of criterion-referenced tests*. Monmouth, OR: Teaching Research, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Hartley, J., & Davies, I. K. Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. *Review of Educational Research*, 1976, *46*, 239-265.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*, 5-10.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project*. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H. L., & Page, S. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

Hsu, T., & Carlson, M. Test constructional aspects of the computer assisted testing model. *Educational Technology*, 1973, *13*(3), 26-27.

Johnson, K. J. Pitt's computer-generated chemistry exam. In *Proceedings of the Conference on Computers in Undergraduate Curricula*, 1973, 199-204.

Keller, F. S. Goodbye, teacher... *Journal of Applied Behavior Analysis*, 1968, *1*, 79-89.

Mager, R. F. *Preparing instructional objectives*. Palo Alto, California: Fearon Publishers, 1962.

Markle, S. M. They teach concepts don't they? *Educational Researcher*, 1975, *4*(6), 3-9.

Markle, S. M., & Tiemann, P. W. Some principles of instructional design at higher cognitive levels. In R. Ulrich, T. Stachnik & T. Mabry (Eds.), *Control of human behavior* (Vol. III). Glenview, Illinois: Scott-Foresman, 1974, pp. 312-323.

McGrath, J. E. A multi-facet approach to classification of individual, group, and organizational concepts. In B. Indik and K. Berrien (Eds.), *People, groups, and organizations: An effective integration*. New York: Teachers College Press, 1967, pp. 191-215.

Melton, R. F. Resolution of conflicting claims concerning the effect of behavioral objectives on student learning. *Review of Educational Research*, 1978, *48*, 291-302.

Miller, H. G., Williams, R. G., & Haladyna, T. M. *Beyond facts: Objective ways to measure thinking*. Englewood Cliffs, N.J.: Educational Technology, 1978.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Company, 1974. (a)

Millman, J. Sampling plans for domain-referenced tests. *Educational Technology*, 1974, *14*, 17-21. (b)

Millman, J., & Outlaw, W. S. *Testing by computer*. Ithaca, New York: Cornell University Extension Publications, 1977.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Olympia, P. L., Jr. Computer generation of truly repeatable examinations. *Educational Technology*, 1975, *14*(6), 53-55.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95-104.

Popham, W. J. *Educational evaluation*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Robin, A. Behavioral instruction in the college classroom. *Review of Educational Research*, 1976, *46*, 313-354.

Roid, G. H. Selecting CAI author languages to solve instructional problems. *Educational Technology*, 1974, *14*(5), 29-31.

Roid, G. H., & Haladyna, T. M. A comparison of objective-based and modified-Bormuth item writing techniques. *Educational and Psychological Measurement*, 1978, *35*, 19-28.

Seddon, G. M. The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research*, 1978, *48*, 303-323.

Shannon, C. E., & Weaver, W. *The mathematical theory of communication*. Urbana: University of Illinois Press, 1949.

Shoemaker, D. M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-148.

Suppes, P., Jerman, M., & Groen, G. J. Arithmetic drills and review on a computer-based teletype. *Arithmetic Teacher*, 1966, April, 303-308.

Tiemann, P. W., & Markle, S. M. *Analyzing instructional content: A guide to instruction and evaluation*. Champaign, IL.: Stipes Publ., 1978.

Vickers, F. D. Creative test generators. *Educational Technology*, 1973, *13*(3), 43-44.

Williams, R. G. A behavioral typology of educational objectives for the cognitive domain. *Educational Technology*, 1977, *17*(6), 39-46.

Williams, R. G., & Miller, H. G. Constructing higher level multiple-choice questions covering factual content. *Educational Technology*, 1973, *13*(5), 39-42.

1117

REFERENCE NOTES

1. Shavelson, R. J. *A method for examining subject-matter structure in written material*. Paper presented at the annual meeting of the American Psychological Association, New Orleans, September 1974.

2. Haladyna, T., & Roid, G. R. *An empirical comparison of three approaches to achievement testing*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1977.

3. Berk, R. A. *An application of structural facet theory to objective based achievement testing*. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, Virginia, March 1978.

4. Ellis, J. A., Wulfeck, II, W. H., Merrill, M. D., Richards, R. E., Schmidt, R. V., & Wood, N. D. *Interim training manual for the instructional quality inventory (IQI)* (NPRDC Tech. Note 78-5). San Diego: Navy Personnel Research and Development Center and COURSEWARE, Inc., February 1978.

5. Finn, P. J. *Generating domain-referenced, multiple-choice test items from prose passages*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

6. Roid, G., & Haladyna, T. *A comparison of several linguistic-based, multiple-choice item writing algorithms*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

7. Roid, G. H., & Finn, P. J. *Algorithms for developing test questions from sentences in instructional materials* (NPRDC Tech. Rep. 78-23). San Diego: Navy Personnel Research and Development Center, 1978.

8. Finn, P. J. *Word frequency, information theory and Cloze performance: A lexical-marker, transfer-feature theory of processing in reading*. Unpublished paper, State University of New York at Buffalo, School of Education, 1977.

9. Roid, G., Haladyna, R., Shaughnessy, J., & Finn, P. *Item writing for domain-based tests of prose learning*. Manuscript submitted for publication, 1978.

10. Engel, J. D., & Martuza, V. R. *A systematic approach to the construction of domain-referenced multiple-choice test items*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., September 1976.

11. Tiemann, P., Kroeker, L. P., & Markle, S. M. Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

12. Tiemann, P. W., & Markle, S. M. Domain-referenced testing in conceptual learning. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

13. Braby, R., Parrish, W. F., Guitard, C. R., & Aagard, J. A. Computer-aided authoring of programmed instruction for teaching symbol recognition (TAEG Report No. 58). Orlando, Florida: Training Analysis and Evaluation Group, 1978.

SECTION 13

TRAINING PROGRAMS AND PROBLEMS

AIRCREW TRAINING RESEARCH - PROJECT ACTIVE

by

Captain Wayne E. Keates
Staff Officer Analysis
Air Command Headquarters
Westwin, Manitoba
Canada R2R 0T0

The views and opinions expressed are those of the author and not necessarily those of the Department of National Defence.

A paper prepared for the 20th Annual Conference of the Military Testing Association held at Oklahoma City, Oklahoma, USA during the week of 30 Oct - 3 Nov 1978.

1121

In early 1976, as part of the on-going development of the new Air Command in the Canadian Forces, an initial analysis of the pilot training system was carried out. It quickly became evident that the training data available was fragmented and that training units were not in a position to collate and analyze the data required to monitor the overall training process. Since Air Command had been tasked with the responsibility for all air-related training, it was proposed that a centralized information system be established. This system would initially include only pilot and navigator training; but after sufficient time had passed to evaluate the system, consideration would be given to establishing a similar monitoring mechanism for all air-related training.

This centralized information system was named the Air Command Training Information System for Validation/Evaluation (Project ACTIVE). Basically it is a longitudinal data collection process in which Air Command training units forward relevant information to a central office in Air Command Headquarters where it can be used for the management of training. In addition, it is proposed that much of the data will be fed back to the complete training system, after some reduction and analysis. In this way, training units will have access to a considerable amount of data which had earlier been unavailable to them and will be better able to consider their own part within the total training process.

At all stages of training we collect three types of information.

1. The first type may be considered biographical. It is used to identify and describe the student. It includes information such as age, entry plan, previous flying experience, etc.

2. The second type is performance information. This merely records how well the student has performed on each course. At more advanced training levels this will also include an assessment of how well prepared the student was for the current course.

3. The third type is attitudinal information. For this, a 43-item training satisfaction questionnaire is administered at each stage of training. Some of the more biographical data is also related to attitudinal variables, such as student's future employment preferences and military assessment (adaptation to military life).

As was mentioned, both pilot and navigator training are included in Project ACTIVE. A brief look at each should make the data flow more obvious.

Figure 1 depicts pilot progression through training to employment. All pilot trainees complete Basic Officer Training, Primary Flying Training and Basic Flying Training. Students for Basic Rotary Wing Training are streamed from Basic Flying Training after hour 140, the remainder continue to 200 hours and are sent to either high performance or multi-engine, with a small number being retained in training employment.

Figure 2 depicts the progression of one student through training, showing the information collected at each stage. Since the student is not under Air Command control during his Basic Officer Training, detailed data collection is not possible at that stage. However, his course grade is picked up later from his training records.

Primary Flying Training is a 27-hour course (about 7 weeks including ground school) on the CT134 Musketeer held at Canadian Forces Base Portage La Prairie. At this stage we collect biographical information, grade, performance ratings, academic marks, officer development ratings, military assessment and attitude questionnaire results. The primary reporting form is the modified CF377, shown as Figure A1 in Annex A. The military assessment is a 5-point rating of the student's adaptation to military life. The attitude questionnaire is administered shortly after the student's first solo. It should be mentioned that the attitude questionnaire, at each stage, is administered by the Base Personnel Selection Officer and forwarded directly to Air Command Headquarters. At no time are individual results made available to any training staff.

Continuing with Figure 2, one can see that the student then proceeds to the Basic Flying Course at Canadian Forces Base Moose Jaw. This is a 200-hour course (about 11 months) on the CT114 Tutor. As was mentioned earlier, rotary wing candidates continue to their next course after 140 hours. At the end of this course the student graduates with his pilot's wings and, if a cadet, receives his commission. The primary reporting form, included as Figure A2, includes biographical information, grade, academic average, and military assessment. The attitude questionnaire is administered twice, just after solo and a few months before the end of the course. In addition, we receive the student's progress book, which includes detailed particulars about each flight and trainer session.

In the particular example shown, the student then proceeds to the Basic Fighter Course at Canadian Forces Base Cold Lake. Three details about the reporting form (Figure A3) should be mentioned. One, this is the first of the feedback forms. The rating categories shown on page 2 of Figure A3 consist of skills the student should have developed on his previous course. The staff must rate the student's ability to perform these groups of tasks relevant to the standard which they expect of an incoming student. Two, page 3 of the form provides a mechanism whereby the staff can specify the particular tasks, within these categories, on which the student was especially good or was deficient. The task card for these categories is included as Figure A4. Three, on page 4 the training unit must assess the standard of performance and rate of progress on this course. The category list will differ for each unit and each aircraft type. These categories were determined by the schools before data collection began and are added to the form by the schools.

Data are collected from the Operational Training Units (OTUs) on a similar 4-page form. In each case the categories in Part I (the feedback categories) are the relevant categories for the preceding course and the Part II categories are the ones specified by that OTU.

We have not begun data collection from the operational squadrons but it is expected that the form will be similar to the feedback portion of the OTU form.

Data collection for the multi-engine and rotary wing streams are similar to the example given.

For navigator training the same procedures apply. Navigator progression is shown in Figure 3 and an example of one student's progression is shown in Figure 4. The student begins his navigator training at the Canadian Forces Air Navigation School at Canadian Forces Base Winnipeg. At the end of this course he receives his wings and proceeds to one of the Operational Training Units. The reporting forms are attached as Annex B. As with the pilots, the attitude questionnaire is administered at each stage of training.

At this stage of its development Project ACTIVE has not provided sufficient numbers of records to justify statistical analysis. This situation is a normal condition in any research involving longitudinal tracking. The large number of possible combinations and permutations of the results are obvious.

A few more general points about the project should be mentioned before closing. Project ACTIVE will allow this Headquarters to monitor pilot and navigator training as a total process involving a number of distinct stages rather than as a series of discrete training courses. At the same time, this does not preclude the option of also examining specific courses in isolation.

The information collected will be summarized and reported, not only to the Headquarters officers responsible for training, but also to all of the schools. Thus the schools will be in a better position to participate in decision making. Attitudinal information will be available to the schools, but only in aggregate form to maintain confidentiality for the students. The informal and often haphazard feedback network between schools will be strengthened by the addition of a formal and structured feedback mechanism.

The specific procedures and possible outcomes that make up Project ACTIVE are not new. Most of the procedures have been used on individual courses in the past. What Project ACTIVE does promise is the opportunity to manage and monitor the longitudinal training process in considerable detail to ensure that our training is the best we can make it.

Figure 1

Pilot Progression

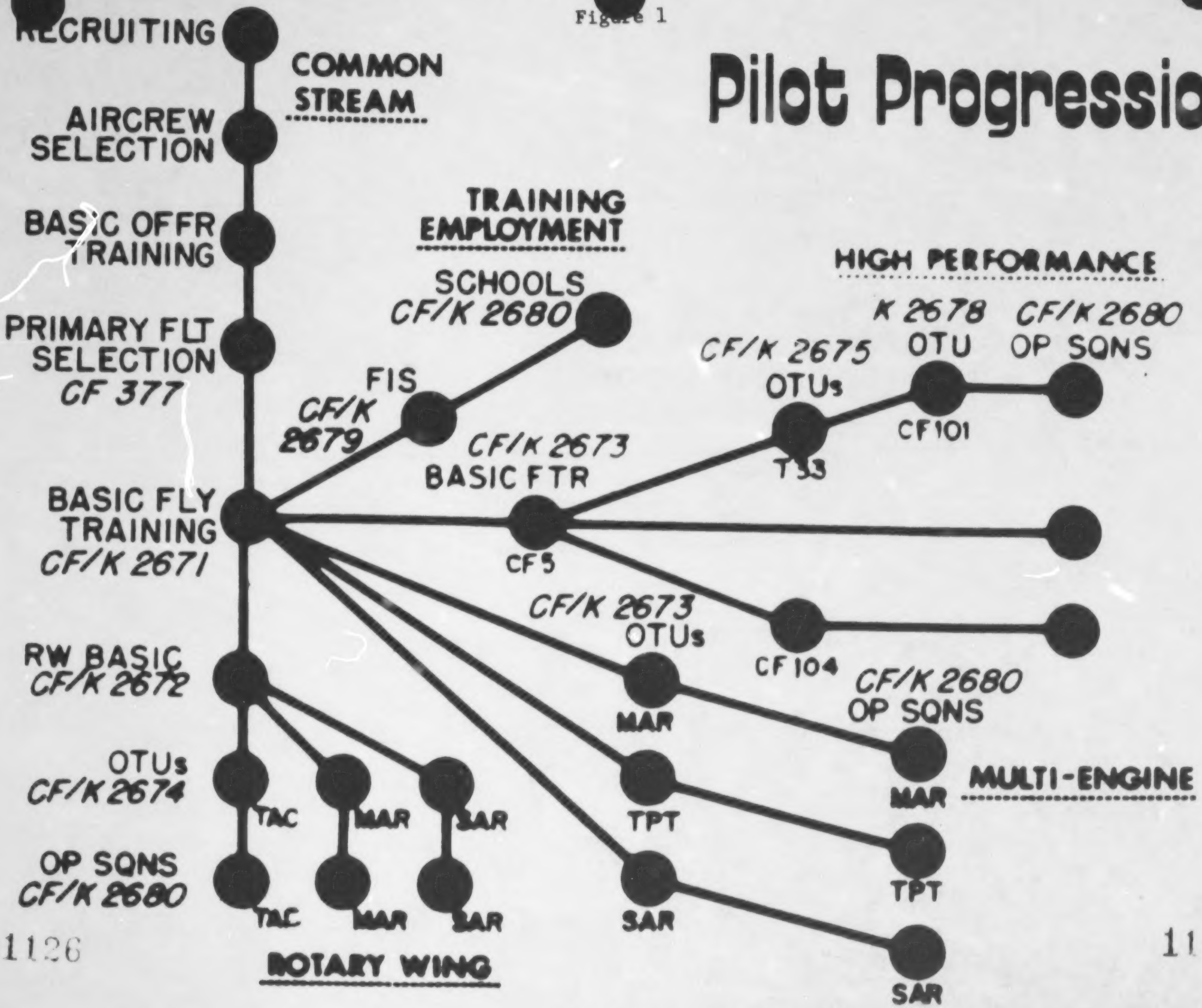
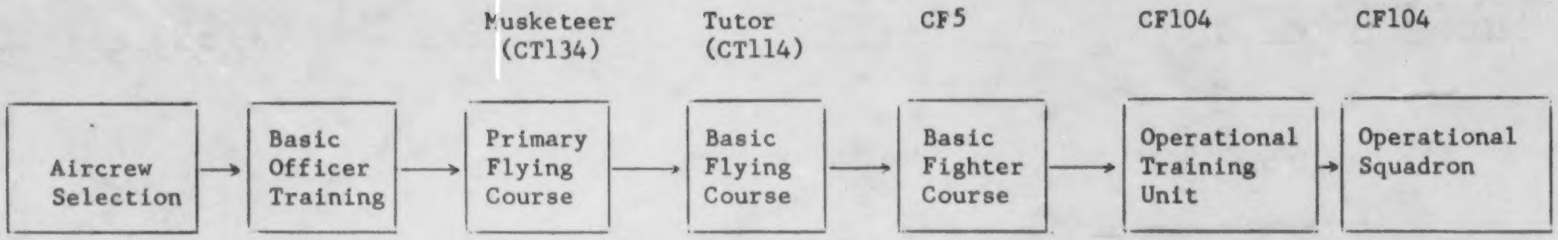

RECRUITING
COMMON STREAM
AIRCREW SELECTION
BASIC OFFR TRAINING
PRIMARY FLT SELECTION CF 377
BASIC FLY TRAINING CF/K 2671
RW BASIC CF/K 2672
OTUs CF/K 2674
OP SQNS CF/K 2680
TAC
MAR
SAR
TAC
MAR
SAR
ROTARY WING
TRAINING EMPLOYMENT
SCHOOLS CF/K 2680
FIS
CF/K 2679
CF/K 2673 BASIC FTR
CF5
HIGH PERFORMANCE
CF/K 2675 OTUs
T33
K 2678 OTU
CF101
CF/K 2680 OP SQNS
CF104
CF/K 2673 OTUs
MAR
TPT
SAR
CF/K 2680 OP SQNS
MAR
TPT
SAR
MULTI-ENGINE

Figure 2

ONE STUDENT'S PROGRESSION

THROUGH PILOT TRAINING


Musketeer (CT134)
Tutor (CT114)
CF5
CF104
CF104
Aircrew Selection
Basic Officer Training
Primary Flying Course
Basic Flying Course
Basic Fighter Course
Operational Training Unit
Operational Squadron

Basic Officer Training	Primary Flying Course	Basic Flying Course	Basic Fighter Course	Operational Training Unit	Operational Squadron
Grade	Biographical	Biographical	Biographical	Biographical	Biographical
	Grade	Grade	Feedback Ratings	Feedback Ratings	Feedback Ratings
	Performance Ratings	Academic Average	Course Ratings	Course Ratings	Military Assessment
	Academic Marks	Military Assessment	Military Assessment	Military Assessment	Attitude Measure
	Officer Development Ratings	Attitude Questionnaire (twice)	Attitude Questionnaire	Attitude Questionnaire	
	Military Assessment	Progress Books			
	Attitude Questionnaire				

Figure 3

Navigator Progression

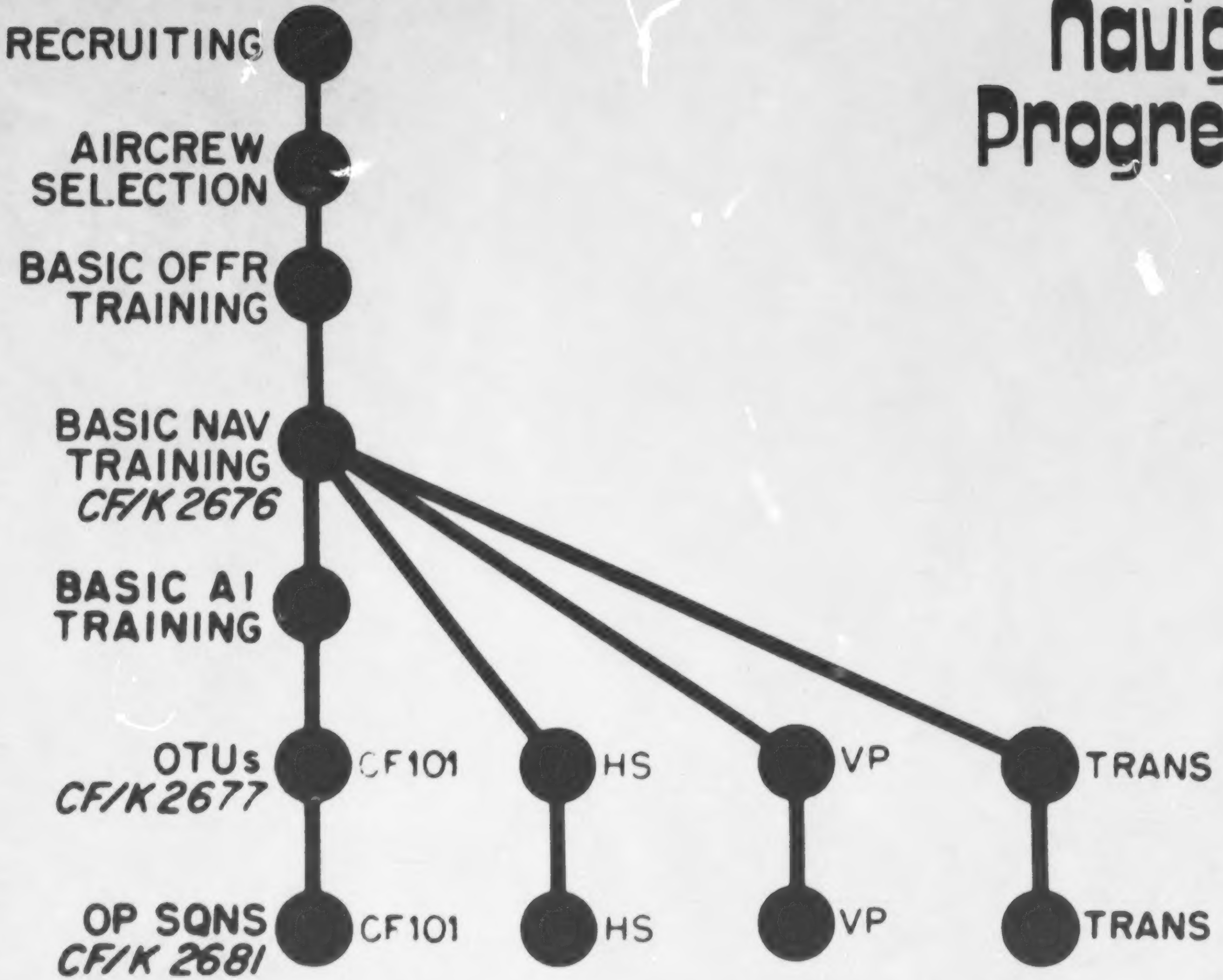
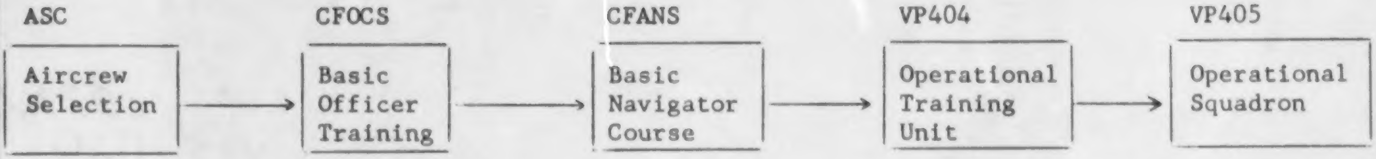

RECRUITING
AIRCREW SELECTION
BASIC OFFR TRAINING
BASIC NAV TRAINING
CF/K 2676
BASIC AI TRAINING
OTUs
CF/K 2677
CF101
HS
VP
TRANS
OP SQNS
CF/K 2681
CF101
HS
VP
TRANS

Figure 4

ONE STUDENT'S PROGRESSION THROUGH NAVIGATOR TRAINING


ASC
Aircrew Selection
CFOCS
Basic Officer Training
CFANS
Basic Navigator Course
VP404
Operational Training Unit
VP405
Operational Squadron

ASC	CFOCS	CFANS	VP404	VP405
	Grade	Biographical	Biographical	Biographical
		Grade	Feedback Ratings	Feedback Ratings
		Academic Average	Course Ratings	Military Assessment
		Course Ratings	Military Assessment	Attitude Questionnaire
		Military Assessment	Attitude Questionnaire	
		Attitude Questionnaire		

CONFIDENTIAL (when completed)
FORMULE CONFIDENTIELLE (une fois remplie)

CANADIAN FORCES
COURSE REPORT

FORCES CANADIENNES
RAPPORT DE COURS

NOTES Instructions for preparation and distribution are contained in CFAO 26-12
NOTA *Les instructions relatives à la façon de remplir et de distribuer le présent document se trouvent dans l'OAFC 26-12*

PERSONAL DATA - *RENSEIGNEMENTS PERSONNELS*

1. SN - *Nº d'Ass Soc*	2. Rank - *Grade*	3. Element - *Élément*	4. Surname - *Nom de famille*	Initials - *Initiales*

5. MOC *Classification de l'emploi militaire*	6. Parent Unit - *Unité d'appartenance*	7. Course title - *Titre du cours*

8. Course Code *Nº de code du cours*	9. Training days - *Jours d'instruction* a. Scheduled *Prévus* ______ b. Required *Exigés* ______	10. Course dates - *Dates du cours* From - *Du* To - *A*	11. Course serial No *Nº de série du cours*

COURSE DATA - *RENSEIGNEMENTS AU SUJET DU COURS*

12. a. Pass Unsuccessful/Course Attended *Succès/Échec/Cours Suivi* ______ b. Grade achieved *Note obtenue* ______ c. Minimum standard *Minimum exigé* ______

13. Recommended as a potential instructor for this course
Recommandé comme instructeur éventuel à ce cours Yes *Oui* ☐ No *Non* ☐ N/A *Ne s'applique pas* ☐

14. General comments (indicate strengths and weaknesses. For additional space use reverse side)
Remarques générales (indiquer les points forts et les points faibles. Si plus d'espace est requis, utiliser le verso de cette feuille).

A. FLYING ASSESSMENT

a. HOURS FLOWN AT 3CFFTS DUAL ☐☐☐ SOLO ☐☐☐ TOTAL ☐☐☐

b. (I) COMPARED TO HISTORICAL FLIGHT SELECTION AVERAGE, STUDENT'S PERFORMANCE WAS: ☐ ABOVE AVERAGE ☐ AVERAGE ☐ BELOW AVERAGE ☐ FAILED

c. PREVIOUS PILOT EXPERIENCE: NO ☐ YES ☐ HOURS ☐☐☐

d. DETAILS OF STUDENT PERFORMANCE

ITEM	STUDENT PROGRESS				COMMENTS
	POOR	SLOW	NORMAL	SUPERIOR	
COORDINATION					
CIRCUITS					
APPROACHES					
LANDINGS					
STALLS					
SPINS					
EMERGENCIES					
FORCED LANDINGS					
AIRMANSHIP					

(ASSESSMENT CONTINUED ON REVERSE)

AUTHENTICATION - *CERTIFICATION*

15. I have read this report - *J'ai lu le présent rapport*

______ Date ______ Student *Élève*

16.

______ Date ______ Reporting Officer *Officier qui a rédigé le rapport*

17. Comment by reviewing officer - *Remarques de l'officier qui a vérifié le rapport.*

1133

______ Date ______ Reviewing Officer *Officier qui a vérifié le rapport* ______ Reporting Officer *Officier qui a rédigé le rapport*

CF 3?? (??-??)
7530-21-???-????

CONFIDENTIAL (when completed)
FORMULE CONFIDENTIELLE (une fois remplie)

CONFIDENTIAL (WHEN COMPLETED)

B. ACADEMIC

STUDENT AVERAGE ______ HISTORICAL COURSE AVERAGE ______

SUBJECT	PASS %	PROGRESS TEST	FINAL	SUPPLEMENTAL
AOI/ENG				
FLIGHT PROCEDURES				
NAVIGATION				
AERODYNAMICS				
INSTRUMENTS				
METEOROLOGY				

C. TECHNICAL VOCABULARY

COMPLETED ☐ NOT REQUIRED ☐

D. OFFICER DEVELOPMENT (Do NOT complete D1 or D2 for cross-trainees.)

CHARACTERISTIC	ASSESSMENT			
	1	2	3	4
1. APPEARANCE & BEARING				
2. CONDUCT				
3. ATTITUDE				
4. INITIATIVE/SELF CONFIDENCE				
5. EFFECTIVENESS UNDER STRESS				
6. ABILITY TO LEARN				
7. PHYSICAL FITNESS/SPORTS				

NOTE - 1 Very Poor 4 Superior

- COMMENTS -

E. MILITARY ASSESSMENT (NOT COMPLETED FOR CROSS-TRAINEES)

HOW WELL HAS THIS OFFICER ADAPTED TO MILITARY LIFE?

NOT WELL ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 VERY WELL

F. AIRSICKNESS (IF APPLICABLE)

HOW OFTEN	ONCE ☐	TWICE ☐	THREE TIMES ☐	MORE THAN THREE ☐
INTERFERENCE WITH PROGRESS	NOT AT ALL ☐	BRIEF ☐	SERIOUS ☐	TRUNCATED LESSONS ☐
CONDITION AT END OF COURSE	IMPROVING ☐	NOT IMPROVING ☐	GETTING WORSE ☐	
CAUSE FACTORS	DISORIENTATION ☐	NERVOUS TENSION ☐	MOTION ☐	
	PHYSIOLOGICAL ☐	OTHER (specify) ☐		

Figure A2

CONFIDENTIAL (WHEN COMPLETED)

AIR COMMAND

TRAINING INFORMATION ▵ VALIDATION/EVALUATION REPORT

INSTRUCTIONS – Complete for each graduate or for students who fail or withdraw. Fill in Item 1 then check appropriate blocks.

1 SIN | NAME | COURSE NO. | RANK: CDT ☐ 2 LT ☐ LT ☐ CAPT ☐ MAJ ☐ LCOL ☐

2a. PLAN: ROTP CMC ☐ CIV ☐ | DEO ☐ | OCTP MIL ☐ CIV ☐ | CROSS TRAINEE ☐ | CFR ☐ | RES ☐ | JDF ☐ | RNLAF MIL COL ☐ RES ☐ X-TRAIN ☐ | FNAT 1 ☐ 2 ☐

2b. AGE: UNDER 18 ☐ 18 ☐ 19 ☐ 20 ☐ 21 ☐ 22 ☐ 23 ☐ 24 ☐ 25 ☐ 26 ☐ 27 ☐ 28 ☐ 29 ☐ 30 ☐ 31 ☐ 32 ☐ 33 ☐ 34 ☐ 35 ☐ OVER 35 ☐

2c. LANGUAGE: ANGLO ☐ FRANCO ☐ OTHER (A) ☐ OTHER (F) ☐ PROFILE ☐☐☐☐

3 BOTC GRADE: A ☐ B ☐ C ☐

WAITING TIMES	UNDER 2 MOS	2-6 MOS	OVER 6 MOS	NOT APPLICABLE
PARU TO BOTC	☐	☐	☐	☐
BOTC TO PORTAGE	☐	☐	☐	☐
PORTAGE TO MOOSE JAW	☐	☐	☐	☐

4a. 2 CFFTS RESULTS: PASS A ☐ B ☐ C ☐ | FAIL ☐ | RECOURSE ☐ | VOLUNTARY WITHDRAWAL ☐

DISPOSITION DATE ☐☐☐☐☐☐ Y Y M M D D

4b. FAILURE REASON: DEFICIENT FLYING SKILLS ☐ MEDICAL ☐ ACADEMIC ☐ CONDUCT OR OFFR DEV ☐

RECOURSE REASON: FLYING ☐ MEDICAL ☐ ACADEMIC ☐ LANGUAGE ☐ OTHER ☐

4c. VOLUNTARY WITHDRAWAL

PROGRESS SATISFACTORY TO VW: YES ☐ NO ☐

HAS POTENTIAL TO GRADUATE: YES ☐ NO ☐

THIS SECTION FOR PSO USE ONLY

DOESN'T LIKE MILITARY LIFE ☐	FEELS INADEQUATE ☐	FEELS PROGRESS UNSATISFACTORY ☐
DOESN'T LIKE FLYING ☐	LANGUAGE ☐	FINANCIAL ☐
FAMILY REASONS ☐	FEELS MORE SUITED TO ANOTHER CLASS ☐	OTHER ☐

5 STUDENT POSTED TO: MULTI-ENGINE SAR ☐ MAR ☐ TPT ☐ | HI-PERF ☐ | INSTR DUTIES ☐ | ROTARY WING SAR ☐ MAR ☐ TAC ☐ | OTHER 1 ☐ 2 ☐

STUDENT'S CHOICE

1. ____________________
2. ____________________
3. ____________________

NO PREFERENCE ☐

6 FLYING HOURS AT 2 CFFTS: DUAL ☐☐☐ SOLO ☐☐☐ TOTAL ☐☐☐

7 ACADEMIC PERFORMANCE

% MARK ☐☐☐

STANDING: TOP ☐ MIDDLE ☐ BOTTOM ☐

MILITARY ASSESSMENT

HOW WELL HAS THIS OFFICER ADAPTED TO MILITARY LIFE?

NOT WELL 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ VERY WELL

CHECK ONE – PIPELINE STUDENTS ONLY

CF/K 2671 (MAR 78)

FORWARD TO – SO ANALYSIS
AIRCOM, CFB WINNIPEG

PROJECT ACTIVE

CONFIDENTIAL (WHEN COMPLETED)

Figure A3

CONFIDENTIAL (WHEN COMPLETED)

AIR COMMAND

SIC ITUR AD ASTRA

Validation/Evaluation Report OTUs

Part I. Wings Validation
Part II. OTU Course Report

GENERAL INSTRUCTIONS

This is the OTU validation and course result form. Part I (Validation) is to be completed when student has demonstrated sufficient performance to assess. Part II (OTU Course Assessment) is to be done when student has completed the course or is CT'd.

PROJECT ACTIVE

SIN	RANK	NAME	
COURSE SERIAL	A/C TYPE	UNIT	UIC

CF/K 2673 MAR 78

FORWARD TO – SO ANALYSIS
AIRCOM, CFB WINNIPEG

CONFIDENTIAL (WHEN COMPLETED)

PART I - WINGS VALIDATION

PROJECT ACTIVE – Form CF/K 2673

This form is the primary validation document for pilot training in the Canadian Forces, and for members of other participating forces. You are asked to comment on the graduate's performance in your OTU based upon the tasks accomplished in Basic Flying Training.

INSTRUCTIONS – PT I

1. You must assess categories of tasks as shown in the list below. Enter rating in the block beside the category. (Omit those which are not exercised in your unit; i.e. Advanced Manoeuvres in M/E OTUs.)

RATING LEGEND

1. Graduate's CATEGORY or TASK performance is:
1. completely unacceptable for this unit;
2. sub-standard for this unit;
3. standard for this unit; or
4. of the highest order.

a. CATEGORY RATINGS

CATEGORY	RATING			
	1	2	3	4
BA. BASIC FLYING SKILLS				
BB. INTERMEDIATE MANOEUVRES				
BC. ADVANCED MANOEUVRES				
BD. BASIC INSTRUMENT MANOEUVRES				
BE. TACAN PROCEDURES				
BF. VOR				
BG. RADAR				
BH. ILS				
BJ. IFR CROSS-COUNTRY PROCEDURES				
BK. AIR NAVIGATION				
BL. NIGHT FLYING (DUALS)				
BM. BASIC FORMATION MANOEUVRES				
BN. INTERMEDIATE FORMATION MANOEUVRES				
BP. ADVANCED FORMATION MANOEUVRES				

b. HOURS FLOWN AT COURSE ENTRY (NON-PIPELINE ONLY)

M/E	RW	HI PERF	TOTAL

c. WAITING TIME - MOOSE JAW to OTU (if applicable) ______ MONTHS

3---

d. TASK RATINGS

You should assess specific tasks if possible. Rate only those tasks that are below or above standard. (1, 2 or 4.) Enter task number and rating in columns below. See Task Card for specific tasks and numbers.

TASK		GRADE	TASK		GRADE	TASK		GRADE	TASK		GRADE

ASSESSED BY

SIN	RANK	SIGNATURE	POSITION	DATE

4 - - -

PART II - OTU COURSE ASSESSMENT

This is a record of the student's accomplishments during this course. The assessment categories are the major phases of your course (eg Weapon Delivery, etc, etc) or the actual objectives in the CTS. Two assessments per category are required. Firstly progress or 'speed of learning' and secondly, the standard assessment as per Rating Legend on page 2.

INSTRUCTIONS

1. Enter all categories in spaces below.
2. Check appropriate column under Progress and Standards.
3. Check appropriate blocks in overall assessment.
4. Complete Military Assessment.
5. Ensure that Attitude Questionnaire is completed.

RATINGS – PROGRESS ASSESSMENT

1. Unacceptable progress
2. Slow Progress
3. Advanced as planned
4. Superior Progress

a. SPECIFIC RATINGS

CATEGORY/OBJECTIVE	PROGRESS				STANDARD			
	1	2	3	4	1	2	3	4

b. OVERALL ASSESSMENT

PROGRESS RATING 1 2 3 4

STANDARDS RATING 1 2 3 4

c. MILITARY ASSESSMENT

HOW WELL HAS THIS OFFICER ADAPTED TO MILITARY LIFE?

NOT WELL				VERY WELL
1	2	3	4	5

CHECK ONE – PIPELINE STUDENTS ONLY

d. DATE OTU COMPLETED

FLYING TIME ON COURSE

e. ADDITIONAL COMMENTS BY SQN COMD

Additional remarks including specific reasons for CT, comments on attitude or officer development and opinion of entry standard, previous coursing, etc. may be appended on separate sheet.

Figure A4

PROJECT ACTIVE TASK/CATEGORY CARD 1

(BASIC FIXED WING - USE WITH FORM CF/K [illegible] 2673 PART 1)

BA - BASIC FLYING SKILLS

B 001	GROUND HANDLING
B 002	TAKE OFF
B 003	CLIMB
B 004	STRAIGHT & LEVEL FLT
B 005	CHANGING AIRSPEED
B 006	GENTLE TURNS
B 007	MEDIUM TURNS
B 009	DESCENTS
B 010	LEVEL-OFF
B 016	TRAFFIC PATTERNS
B 017	CIRCUIT (NORMAL)
B 018	OVERSHOOT
B 019	LANDING (BASIC)

BB - INTERMEDIATE MANOEUVRES

B 020	STRAIGHT-IN LANDING
B 008	STEEP TURNS
B 011	SLOW FLYING
B 012	LANDING ATTITUDE STALL
B 013	FINAL-TURN STALLS
B 014	HIGH-SPEED STALLS
B 015	UNUSUAL ATTITUDES
B 021	CLOSED PATTERNS
B 022	FORCED LANDING
B 023	FLAPLESS LANDING
B 024	FORCED LANDING FROM TP
B 025	SPINS
B 026	RANDOM RADAR
B 027	MINIMUM-ROLL LANDING
B 028	SQUARE CIRCUIT
B 030	SLOW ROLL
B 031	LOOP (BELOW 30,000')
B 032	MAXIMUM-RATE TURNS
B 033	CLOVER LEAF
B 034	CUBAN EIGHT
B 037	BARREL ROLL
B 038	FOUR-POINT ROLL
B 039	HESITATION ROLL
B 040	ROLL-OFF-THE-TOP
B 041	HALF-ROLL & PULL-THRU
B 046	EMERGENCY DESCENT
B 049	ROLL-IN AND ROLL-OUT

BC - ADVANCED MANOEUVRES

B 029	UNUSUAL ATTITUDE SPIN
B 035	OFF-SPEED AEROBATICS
B 036	LOW FLYING
B 042	MULTIPLE AEROBATICS
B 043	MACH RUN
B 044	LOOP (ABOVE 30,000')
B 045	ROLL (ABOVE 30,000')
B 047	VERTICAL EIGHT
B 048	VERTICAL ROLL

BD - BASIC INSTRUMENT MANOEUVRES

B 060	STRAIGHT-AND-LEVEL FLIGHT
B 061	CHANGING AIRSPEED
B 062	GENTLE TURNS
B 063	MEDIUM TURNS
B 064	CLIMBS
B 065	DESCENTS
B 066	LEVEL-OFF
B 067	TURNS TO HEADING
B 068	RATED CLIMB
B 069	RATED DESCENT
B 070	STANDARD-RATE TURNS
B 071	STEEP TURNS
B 072	UNUSUAL ATTITUDE
B 073	TIMED TURNS
B 078	UHF HOMING
B 081	CLEARANCES (LOCAL)
B 091	TIMED TURN (NO DRMI)
B 092	UNUSUAL ATTITUDES

BE - TACAN PROCEDURES

B 074	TACAN INTERCEPTION
B 075	TACAN TRACKING
B 076	TACAN DEPARTURE
B 077	TACAN ARCING
B 079	TACAN POINT-TO-POINT
B 080	TACAN HOLDING
B 082	TACAN APPROACH
B 083	TACAN MISSED APPROACH
B 084	MIN FUEL TACAN APPROACH

BF - RADAR

B 086	RANDOM RADAR
B 085	LANDING FROM RADAR
B 087	RADAR FINAL
B 088	RADAR SQUARE PATTERN
B 089	RADAR MISSED APPROACH
B 090	NO-COMPASS RADAR

BG - VOR

B 093	VOR INTERCEPTION
B 094	VOR TRACKING
B 095	VOR HOLDING
B 096	VOR APPROACH
B 097	LANDING FROM VOR APPROACH
B 098	VOR MISSED APPROACH

BH - ILS

B 099	RADAR VECTORED ILS
B 100	ILS BACK CRSE FINAL
B 101	ILS FRONT CRSE FINAL
B 102	LANDING FROM ILS APPROACH
B 103	ILS MISSED APPROACH
B 104	TACAN/ILS APPROACH

SEE OVER ....

BJ - IFR X-CTRY

B 105	PRE-FLIGHT PLANNING (X-CTRY)
B 106	IFR CLEARANCES
B 107	DEPARTURE
B 108	ENROUTE
B 109	TACAN APPROACH
B 110	ENROUTE RADAR DESCENT
B 111	RADAR FINAL APPROACH
B 112	VOR APPROACH
B 113	VOR/ILS APPROACH
B 114	TACAN/ILS APPROACH
B 115	TACAN/RADAR PICK-OFF
B 116	RADAR VECTORED ILS APPROACH
B 117	ILS BACK CRSE FINAL APPROACH
B 118	ILS FRONT CRSE FINAL APPROACH
B 119	LANDING FROM INSTRUMENT APPROACH
B 120	INSTRUMENT MISSED APPROACH
B 121	VISUAL APPROACH AND LANDING

BL - NIGHT FLYING

B 140	GROUND HANDLING
B 141	TAKE-OFF
B 142	CLIMB
B 143	UNUSUAL ATTITUDE RECOVERY
B 144	TRAFFIC PATTERN
B 145	CIRCUIT
B 146	LANDING
B 147	OVERSHOOT
B 148	TACAN APPROACH
B 149	LANDING FROM INSTRUMENT APPROACH
B 150	RADAR APPROACH
B 151	MISSED APPROACH

BK - AIR NAVIGATION

B 130	PREPARATION AND FLIGHT PLANNING
B 131	MEDIUM-LEVEL NAVIGATION PROCEDURES
B 132	MAP READING
B 133	SET HEADING PROCEDURES
B 134	LOG KEEPING AND ENTRIES
B 135	PILOT ABILITY AND AIRMANSHIP
B 136	TAKE-OFF AND DEPARTURE
B 137	MEDIUM-LEVEL NAVIGATION PROCEDURES (MDR)
B 138	LOW-LEVEL NAVIGATION PROCEDURES (BASIC)

BM - BASIC FORMATION MANOEUVRES

B 160	GROUND HANDLING
B 161	STATION KEEPING (TO 45°)
B 162	CHANGING STATION
B 163	WING TAKE-OFF
B 164	SELECTION OF ANCILLARIES
B 165	REJOIN
B 166	TRAIL
B 167	FLAT TURNS (4 PLANE)
B 168	GROUND HANDLING (4 PLANE)
B 169	STATION KEEPING (4 PLANE)
B 170	CHANGING STATION (4 PLANE)

BN - INTERMEDIATE FORMATION MANOEUVRES

B 171	WING LET-DOWN
B 172	TRAFFIC PATTERN
B 173	CIRCUIT
B 174	LANDING
B 175	STATION KEEPING (OVER 45°)
B 176	MISSED APPROACH
B 177	INTERVAL TAKE-OFF (4 PLANE)
B 178	JOIN-UP (4 PLANE)
B 179	TRAFFIC PATTERN (4 PLANE)
B 180	CIRCUIT (4 PLANE)
B 181	LANDING (4 PLANE)

BP - ADVANCED FORMATION MANOEUVRES

B 182	LEADING
B 183	INSTRUMENT APPROACH
B 184	FORMATION LANDING

Figure B1

ANNEX B

CONFIDENTIAL (WHEN COMPLETED)

AIR COMMAND

TRAINING INFORMATION △ VALIDATION/EVALUATION REPORT

INSTRUCTIONS – Complete for each graduate or for students who fail or withdraw. Fill in Item 1 then check appropriate blocks.

1 SIN | NAME | COURSE NO. | RANK CDT ☐ 2 LT ☐ LT ☐ CAPT ☐ MAJ ☐ LCOL ☐

2a. PLAN ROTP CMC ☐ CIV ☐ DEO ☐ OCTP MIL ☐ CIV ☐ CROSS TRAINEE ☐ CFR ☐ RES ☐ FNAT 1 ☐ FNAT 2 ☐

2b. AGE UNDER 18 ☐ 18 ☐ 19 ☐ 20 ☐ 21 ☐ 22 ☐ 23 ☐ 24 ☐ 25 ☐ 26 ☐ 27 ☐ 28 ☐ 29 ☐ 30 ☐ 31 ☐ 32 ☐ 33 ☐ 34 ☐ 35 ☐ OVER 35 ☐

2c. LANGUAGE ANGLO ☐ FRANCO ☐ OTHER (A) ☐ OTHER (F) ☐ PROFILE ☐☐☐☐

3 BOTC GRADE A B C

WAITING TIMES	UNDER 2 MOS	2-6 MOS	OVER 6 MOS	NOT APPLICABLE
PARU TO BOTC	☐	☐	☐	☐
BOTC TO CFANS	☐	☐	☐	☐

4a. CFANS RESULTS PASS A B C FAIL ☐ RECOURSE ☐ VOLUNTARY WITHDRAWAL ☐

DISPOSITION DATE ☐☐☐☐☐☐ Y Y M M D D

4b. FAILURE REASON DEFICIENT FLYING SKILLS ☐ MEDICAL ☐ ACADEMIC ☐ CONDUCT OR OFFR DEV ☐

RECOURSE REASON FLYING ☐ MEDICAL ☐ ACADEMIC ☐ LANGUAGE ☐ OTHER ☐

4c. VOLUNTARY WITHDRAWAL

PROGRESS SATISFACTORY TO VW YES ☐ NO ☐

HAS POTENTIAL TO GRADUATE YES ☐ NO ☐

THIS SECTION FOR PSO USE ONLY

DOESN'T LIKE MILITARY LIFE ☐ FEELS INADEQUATE ☐ FEELS PROGRESS UNSATISFACTORY ☐

DOESN'T LIKE FLYING ☐ LANGUAGE ☐ FINANCIAL ☐

FAMILY REASONS ☐ FEELS MORE SUITED TO ANOTHER CLASS ☐ OTHER ☐

5 STUDENT POSTED TO VP ☐ HS ☐ CF101 ☐ TRANS ☐ OTHER 1 ☐ 2 ☐

STUDENT'S CHOICE

1. ____________
2. ____________
3. ____________

NO PREFERENCE ☐

6 FLYING HOURS AT CFANS ☐☐☐

7 ACADEMIC PERFORMANCE % MARK ☐☐☐

STANDING TOP ☐ MIDDLE ☐ BOTTOM ☐

MILITARY ASSESSMENT

HOW WELL HAS THIS OFFICER ADAPTED TO MILITARY LIFE?

NOT WELL 1 2 3 4 5 VERY WELL

CHECK ONE – PIPELINE STUDENTS ONLY

CF/K 2676 (MAR 78)

FORWARD TO – SO ANALYSIS
AIRCOM, CFB WINNIPEG

PROJECT ACTIVE

CONFIDENTIAL (WHEN COMPLETED)

1 / 1

PART II - CATEGORY RATING

PROJECT ACTIVE – FORM 2676

This form is the primary evaluation document for Navigator training in the Canadian Forces, and for members of other participating forces. You are asked to comment on the graduate's performance at CFANS based upon the tasks accomplished in Basic Navigator Training.

INSTRUCTIONS – PT II

1. You must assess categories of tasks as shown in the list below. Enter rating in the block beside the category.

RATING LEGEND

1. Graduate's CATEGORY performance is: 1. Unsatisfactory – FAILED
2. Achieved minimum rating
3. Achieved average rating
4. Achieved good rating

a. CATEGORY RATINGS

CATEGORY	RATING			
	1	2	3	4
A. MAPS, CHARTS & FLIGHT DOCUMENTS				
B. FLIGHT PLANNING				
C. PRE-FLIGHT				
D. FUEL MONITORING				
E. ELECTRONIC FIXING AIDS				
F. CELESTIAL FIXING				
G. MPP PROCEDURES				
H. POSITION COMPUTER USE				
J. GRID NAVIGATION				
K. GYRO NAVIGATION				
L. POST FLIGHT				
M. COMMUNICATIONS				
N. AIRMANSHIP				
P. AIR REGULATIONS				
Q. TASK CO-ORDINATION				
R. ACADEMIC EFFORT				

CONFIDENTIAL (WHEN COMPLETED)

AIR COMMAND

Validation/Evaluation Report OTUs

Part I. Wings Validation
Part II. OTU Course Report

GENERAL INSTRUCTIONS

This is the OTU validation and course result form. Part I (Validation) is to be completed when student has demonstrated sufficient performance to assess. Part II (OTU Course Assessment) is to be done when student has completed the course or is CT d.

PROJECT ACTIVE

SIN	RANK	NAME	
COURSE SERIAL	A/C TYPE	UNIT	UIC

FORWARD TO – SO ANALYSIS
AIRCOM, CFB WINNIPEG

CF/K 2677 MAR 78

CONFIDENTIAL (WHEN COMPLETED)

PART I - WINGS VALIDATION

PROJECT ACTIVE – Form CF/K 2677

This form is the primary validation document for Navigator training in the Canadian Forces, and for members of other participating forces. You are asked to comment on the graduate's performance in your OTU based upon the tasks accomplished in Basic Navigator Training.

INSTRUCTIONS – PT I

1. You must assess categories of tasks as shown in the list below. Enter rating in the block beside the category.

RATING LEGEND

1. CFANS graduate's CATEGORY or TASK performance is:
   1. completely unacceptable for this Unit/Sqn
   2. sub-standard for this Unit/Sqn
   3. standard for this Unit/Sqn
   4. of the highest order.

a. CATEGORY RATINGS

CATEGORY	RATING			
	1	2	3	4
A. MAPS, CHARTS & FLIGHT DOCUMENTS				
B. FLIGHT PLANNING				
C. PRE-FLIGHT				
D. FUEL MONITORING				
E. ELECTRONIC FIXING AIDS				
F. CELESTIAL FIXING				
G. MPP PROCEDURES				
H. POSITION COMPUTER USE				
J. GRID NAVIGATION				
K. GYRO NAVIGATION				
L. POST FLIGHT				
M. COMMUNICATIONS				
N. AIRMANSHIP				
P. AIR REGULATIONS				
Q. TASK CO-ORDINATION				
R. ACADEMIC EFFORT				

b. HOURS FLOWN AT COURSE ENTRY (NON-PIPELINE ONLY)

HS	AI	VP	TPT	TOTAL

c. DATE PART I COMPLETED ☐☐☐☐☐☐

3---

1145

PART II - OTU COURSE ASSESSMENT

This is a record of the student's accomplishments during this course. The assessment categories are the major phases of your course (eg Weapon Delivery, etc, etc) or the actual objectives in the CTS. Two assessments per category are required. Firstly progress or 'speed of learning' and secondly, the standard assessment as per Rating Legend on page 2.

INSTRUCTIONS

1. Enter all categories in spaces below.
2. Check appropriate column under Progress and Standards.
3. Check appropriate blocks in overall assessment.
4. Complete Military Assessment.
5. Ensure that Attitude Questionnaire is completed.

RATINGS – PROGRESS ASSESSMENT

1. Unacceptable progress
2. Slow Progress
3. Advanced as planned
4. Superior Progress

a. SPECIFIC RATINGS

CATEGORY/OBJECTIVE	PROGRESS				STANDARD			
	1	2	3	4	1	2	3	4

b. OVERALL ASSESSMENT

PROGRESS RATING ☐ 1 ☐ 2 ☐ 3 ☐ 4

STANDARDS RATING ☐ 1 ☐ 2 ☐ 3 ☐ 4

c. MILITARY ASSESSMENT

HOW WELL HAS THIS OFFICER ADAPTED TO MILITARY LIFE?

NOT WELL ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 VERY WELL

CHECK ONE – PIPELINE STUDENTS ONLY

d. DATE OTU COMPLETED ☐☐☐☐☐☐

FLYING TIME ON COURSE ☐☐☐

e. ADDITIONAL COMMENTS BY SQN COMD

Additional remarks including specific reasons for CT, comments on attitude or officer development and opinion of entry standard, previous coursing, etc. may be appended on separate sheet.

DEVELOPMENT OF THE ARMY ROTC MANAGEMENT SIMULATION PROGRAM AND INSTRUCTORS' ORIENTATION COURSE

R. A. Dapra and W. Byham
Development Dimensions, Inc.

M. G. Rumsey, A. Castelnovo and R. S. Wellins
U. S. Army Research Institute for the Behavioral and Social Sciences

Paper presented at Military Testing Association's Annual Conference, October 23-27, 1978, Oklahoma City, Oklahoma

DEVELOPMENT OF THE ARMY ROTC MANAGEMENT SIMULATION PROGRAM AND INSTRUCTORS' ORIENTATION COURSE

The Army ROTC Management Simulation Program (MSP) is a modular instructional package which provides ROTC cadets with the opportunity to apply and develop basic management skills in realistic, simulated situations. The accompaning Instructor Orientation Course (IOC) is a self-paced, self-contained, instructional program designed to develop the skills required to effectively teach the MSP. The purpose of this paper is to describe the development, objectives and content of each of these programs. In addition, the results of comprehensive field evaluations of both programs will be reported.

Development of the Management Simulation Program

The MSP was conceptualized as a program which would provide skills in the interpersonal and management areas underlying effective leadership. The first step in the development process was determining the management skills to be included in the program. After extensive literature reviews and interviews with managers, several broad, focal management skills were identified and then classified into four separate modules. The first module deals with problem analysis and decision-making; Module II is concerned with management planning and organizing skills; the third module concentrates on management delegation and control; and finally, the fourth module includes instruction in the interpersonal skills required for effective management.

Each of the focal management skills were further divided into a number of instructional units called essential elements. For example, the essential elements for problem analysis are: defining the problem as it relates to your function and goal, collecting and evaluating the facts, determining the relationship between the facts and the problem, and identifying the most likely cause of the problem.

The next step in the developmental process was the establishment of the instructional components for the program. Decisions had to be made on how to best present the material to the ROTC cadet. The most unique aspect of the MSP was the inclusion of specially designed simulations based on assessment center technology. The assessment center method utilizes a series of simulations which are designed to elicit behavior which will actually be required for a given job. In the past, assessment center technology has been used for evaluating management potential. The MSP is unique in that it incorporates assessment center simulations into the educational process, thereby allowing students the opportunity to actively participate in the learning process. Two types of simulation exercises were developed for the MSP. The first type of exercise was designed to elicit and illustrate behaviors related to specific essential elements underlying a particular management skill. The second type of simulation was designed to elicit behaviors underlying all of the essential elements for one or more of the focal management skills. The later type of simulation was called a "capstone" exercise.

Some of the simulations contained in the MSP include the Registrar's Office Fact-Finding Exercise, Organizing and Planning the Bicentennial Exercise; Delegation and Control In-Basket, Executive Director of the Community Fund, etc. Each of these exercises requires a large degree of active student participation. For example, the In-Basket, the capstone exercise for Module III, requires students to play the role of a plant manager. He/she must handle accumulated letters, notes and requests found in a simulated in-basket. The In-Basket contains a total of 20 items which require the student to effectively utilize the essential elements of delegation and control.

You may be wondering why civilian-management settings were selected for use with ROTC cadets. The major factor in this decision was that ROTC cadets are more familiar with the civilian management environment than they are with the Army environment although the skills underlying the functions and responsibilities of both environment are the same. It was felt that unfamiliar problem environments would distract the student from learning the targeted management skills.

Although the simulations were the primary vehicle for instruction, an integrated system linking the simulations with one another and with the essential elements had to be developed and incorporated into the MSP. Consequently, other program components were developed including:

1. Relevant text material which precisely defines the nature of each management skill.

2. Brief lectures which introduce each module and illustrate the management skills.

3. Group presentation and/or discussion of the results of each simulation.

4. Highly structured feedback and reinforcement of appropriate responses relative to each exercise and the specific essential elements.

The complete MSP consists of four instructional modules containing the simulation exercises and textual material; a videotape interview for use in teaching cadets interpersonal skills; an instructor's workbook; and a booklet containing student evaluation material. Each module is a separate unit so that the instructor has the option of teaching all four modules or any combination of one or more modules.

Evaluation of the Management Simulation Program

Since the MSP is a unique, educational approach in many respects, it was important to determine if the program was practical and feasible for classroom use. Another equally important issue was whether or not the MSP could generate and maintain student interst and involvement. To answer these questions, 21 ROTC programs participated in a comprehensive field evaluation. The instructors teaching the MSP as well as the cadets enrolled in the program were asked to complete a survey designed to gauge cadet and instructor reactions to the program. The quantitative results of this evaluation are too lengthy to review within the time limits of this presentation. However, the following conclusions were reached based on the survey data:

1. The simulation program was generally viewed by both instructors and cadets as effective and interesting.

2. Student materials were generally found to be clear and complete.

3. Instructor material was found to be adequate.

4. In general, the length of the exercises was satisfactory to both instructor and cadet.

Although the overall evaluation of the program was favorable, a few deficiencies and suggestions for improvement surfaced. Based on these suggestions, changes were made to add to the clarity and/or comprehensiveness of the MSP.

Development of the Instructor Orientation Course

One of the suggestions made by instructors in the MSP evaluation was the need for training prior to teaching the MSP. Before the evaluation, 20 ROTC instructors met at a central location to receive guidance on how to teach the MSP. Obviously, centralized training for all prospective MSP instructors would be a costly and impractical venture. An instructor's training program had to be developed which was cost-effective, easy to administer and minimally time consuming for instructors.

In order to meet these needs, an Instructor Orientation Course (IOC) was developed. The IOC was designed to accomplish two major objectives:

1. To provide an opportunity for potential instructors of the MSP to experience the program from a student perspective by actively responding to each exercise and all the other program materials.

2. To provide an opportunity to develop critical instructor competencies relative to each component in the MSP by providing instructional models and/or skill practice relative to each competency.

The first step of the development process concentrated on the identification of the competencies required to effectively teach the MSP. Two workshops were attended by prospective instructors and the MSP developer. During the workshops, the developer presented each component of the MSP to the prospective instructors and required their participation as students. Once the student perspective was achieved, the participants discussed the instructional skills necessary to teach a specific component of the program. Considerable time was spent identifying the types of activities and instruments which would best develop the essential skills required for effective instruction. The information generated in these workshops was analyzed to identify the critical instructor compentencies to be addressed in the IOC. Once identified, the development of an audio-tape, workbook system began.

The IOC consists of three major components. First, a short videotape was developed to provide a detailed introduction to the MSP, encourage the use of the MSP, and provide a review of the various MSP and IOC program components.

The second major component is a series of four audiotapes; one for each of the MSP modules. The audiotapes help the prospective instructor in several ways:

1. They provide an overview to each module.

2. They provide specific directions for responding to the student materials.

3. They review and discuss the objectives of the exercises in the student modules and illustrate typical student responses.

4. They delineate and discuss critical instructor competencies.

5. They discuss and critique competency development activity.

6. They clarify the role of the instructor in administering the student activities.

Lastly, a set of instructor workbooks were developed to be used in conjunction with the audiotapes. It was determined that only Modules I, II and II required workbooks. This decision was predicated on the fact that the student materials for Module IV were already in a format which could be used effectively with the audiotape. The evaluation materials were unique and were already adequately addressed in the Instructor's Manual. The instructor's workbooks were designed to accomplish the following objectives:

1. To provide the student materials for the instructor to study.

2. To provide a format for the instructor to respond to student materials.

3. To provide an opportunity for self-evaluation of responses to student materials.

4. To provide activities to aid in the development of instructor competencies.

5. To provide an opportunity for self-evaluation of the competency development activities.

Instructor Orientation Course Evaluation

Since the MSP represented an innovative and relatively complex instructional approach, the IOC had to overcome resistance to a new mode of instruction, provide a student perspective of the program materials and requirements, and develop instructor competencies for teaching each of the major program components. Consequently, a three-phase evaluation was designed and implemented. The stages included a small scale developer's test, a telephone interview with eight ROTC instructors and a mail questionnaire sent to 16 ROTC instructors. Again, the detailed results are too lengthy to review here. In general, however, the IOC was well received. As one instructor stated: "I enjoyed participating in the program. It was a valuable refresher as well as a new challenge."

Results of the mail questionnaire revealed that instructors found the videotape informative, interesting, and of high quality. After seeing the tape, most instructors highly recommended that the MSP be incorporated into their ROTC curriculum. The audiotapes were rated as effective in preparing instructors to teach the MSP. Lastly, the workshops were evaluated as useful in developing the teaching competencies required to teach the MSP.

The evaluation was also useful in identifying some of the program deficiencies before it became fully operational. The course materials had to be reorganized into one package and a reference list on management was developed and included as part of the course. In addition, the video-tape was shortened and filmed over with a professional narrator. Other minor changes aimed at clarifying the instructions and material were also instituted.

Conclusions.

The Army Field Manual, FM 22-100, defines management as "the process of planning, organizing, coordinating, directing and controlling resources such as men, materials, time and money to accomplish the organizational mission." The MSP was designed with this purpose in mind; to prepare new lieutenants for their jobs as leaders and managers.

The MSP and IOC are novel in that they apply assessment center technology to the educational process. The simulation exercises spark student interest in material that otherwise might be somewhat dry. Evaluation of both programs indicate that this method of instruction is highly stimulating and effective. The evaluation also allowed program deficiencies to be corrected before the final version was produced. The MSP and IOC have recently been distributed to all ROTC regions through TRADOC. It is expected that it will be widely used by host institutions this coming year.

"HOW DO YOU BUY 'GOOD DESIGN'": AN EXAMINATION OF THE ARMY'S TEC PROGRAM

For Presentation at the 20th Annual Conference of the Military Testing Association 30 October-3 November 1978

CPT Robert R. Begland
Training Developments Institute
TRADOC U.S. Army

The Army, in terms of training hardware and software systems certainly could be described as the world's largest buyer of training products. In terms of only training products a conservative estimate of expenditures for 1978 would be approximately $40,000,000.

The dilemma facing any customer is how do you get the best product for the least amount of money. Specifically, in the training arena, the question boils down to, "How does an organization buy 'good design' in its training materials?"

In the areas of equipment technology, the answer is simple: design specifications. But in the area of training technology do we presently have and are we using appropriate design specifications?

The purpose of this presentation is to suggest potential inroads to identifying design considerations that may be appropriate, to describe them in terms of the Army's Training Extension Course (TEC) Program, to examine their research basis, and to offer recommendations on how to insure that good design can be obtained if it can be recognized.

It is essential to state that this presentation is restricting its focus to only the design phase of training development. Acknowledging, as a given, that the "What to train" decision has been made in accordance with a systematic process.

A paraphrased restatement then of the original question is: "How do you buy good design in your training materials," and equally important "how do you recognize it?"

To establish a reference point, "good design" will be loosely defined as: the utilization of the "best" learning theories during the design of instructional materials for a given set of instructional objectives, for a specified target population, and will be evaluated in terms of efficiency and effectiveness.

TEC History

The U.S. Army subsequent to the Vietnam War was able to examine its training programs, and found that there was an acknowledged training deficiency in terms of individual competencies and unit proficiency. The Training Extension Course (TEC) Program was developed as one attempt to reduce this deficiency.

The scope of TEC is mind boggling. Since 1973 1,050 lessons have been produced, 3000 lessons are in various stages of development, and 950 are projected to be developed each year. The total expenditure for the TEC program to date is over 120 million dollars.

TEC utilizes a multimedia format and capitalizes on the subject matter expertise found in the various service schools to produce exportable training material capable of being used by the individual soldier, both active duty and reserve components. The vast majority of these TEC lessons are in an audiovisual format, utilizing a closed-loop filmstrip and an audio cassette, which are played on a Besseler Cue-See projector.

The TEC objectives were:

> INDIVIDUAL SOLDIER - To provide packaged, validated self-administered, individualized, and self-paced instructional materials to soldiers in units to teach those necessary tasks for job/duty proficiency required in both peacetime and combat environments.
>
> SMALL UNIT COMMANDER - To assist the commander in reducing burdensome material preparation time and resources by dedicating training methods and resources to the optimal use of the TEC materials which facilitate the small unit commander's role as a training manager. (TEC, 1975)

Presently, audiovisual materials have emerged as the primary media for TEC. They offer replicability of systematically designed instruction, ease of dissemination, adaptibility to self-paced, individualized, criterion-referenced instruction, and an attractiveness to the target population. Yet what are the design guidelines that are driving this system?

Since the TEC program is task specific, self-paced, and performance oriented, it presupposes learning, will occur because of the subject's interaction with these materials. A recent study observed that, "when the student is considered as an active agent in his own learning, it becomes necessary to emphasize those student activities and processes which give rise to learning."

(Bertou, Clasen, & Lambert, 1972)

The field of study that deals with the manipulation of events and activities within instruction to give rise to learning was labeled "mathemagenics" by Ernest Rothkopf (1963).

This concept is precisely the focus of my research: what are those instructional events that if present in an audio visual lesson will "give rise to learning?"

The vast majority of the research on the mathemagenic theory has been conducted with printed text. Little research has been conducted in the area of audio-visual instruction utilizing the mathemagenics concepts.

The Army, since the elimination of the draft has experienced a change in the ability levels of soldiers now entering. In preparing its instructional materials, technical manuals, job performance aids, and soldier related publications it must accommodate the mental processing abilities of its soldiers. The Army must insure that each of its instructional programs can produce the optimal amount of learning, given a wide variance in degrees of soldier ability.

If the Army is getting a soldier with a lower ability level, perhaps these individuals, labeled as having lower ability, actually just don't have those finite learning skills of attention, perceptual processing, association, abstraction, and encoding which are important during the information presentation phase of instruction.

There are several possible strategies that could be adopted to accommodate the soldier. An audiovisual format is a means of circumventing potential reading disabilities. Highlighting or emphasizing important points with arrows may compensate for poor attention. Representation or repeated viewings of a lesson may affect retention. Providing an advance organizer may set the structure and facilitate acquistion. Questioning strategies certainly have some obvious functions: review questions in their recall of essential prerequisite skills, preadjunctive questions as a focusing instrument for selective attention, and postadjunctive questions as an attention, maintenance function and general search strategy. Feedback has traditionally been an essential component of instruction, along with practice exercise and some type of a self-evaluation. But what direction has research suggested that we go?

If we looked at printed materials it seems that, "It is what the student does with the words he reads while he reads them that determines the efficiency of learning" (Frase, 1968). Appending a corollary to this theorem: It may well be, what the student does with the information presented (orally, visually, or problematically) while engaged in the learning process that determines the effectiveness of learning.

Some research on representation of lessons offers some intriguing findings: the first semantic encoding by a learner is relatively stable over time in spite of representation of the material, and successive retesting over the material. It was observed that:

> subjects are unable to profit so much as one might expect them to from the opportunities for improvement and for making corrections that appear to be provided by the repeated presentation of the material. The version that an individual has himself reproduced appears to be particularly stable in his memory, and hence resistent to changes in the direction either of increased accuracy or increased forgetting. (Howe, 1977)

Another research area that has offered some interesting insights deals with the use of feedback. Feedback can be described as the information presented to the subject immediately after a response to a question that enables the subject to judge the correctness or completeness of that response. Some generalizations have emerged:

1) Feedback is not important if the student has made a correct response to the question in instructional materials.

2) Feedback is critical if the student has made an incorrect response.

3) Feedback is appropriate only when the student has made a faulty interpretation of the materials or question, it is not appropriate for a lack of understanding.

4) If the feedback is readily available it will have no effect.

5) The delayed presentation of feedback for a day increases retention and performance (Kulhavy, 1977)

Undoubtedly, there are appropriate and inappropriate strategies for the design of audio-visual lessons like TEC. Allen (1975) has tried to generalize from research, "What can the designer and producer do to manipulate, arrange, emphasize, or enhance the way the message is presented to optimize learning from it." He concluded that ..."it would appear, therefore, that both empirical evidence and theory point to greater benefit for lower ability learners from procedures that give direction to their inspectional behavior of the instructional stimuli to which they are exposed. Such techniques would be expected to compensate for their poor attentional and discriminational abilities." (Allen, 1975)

The research over the last 15 years has shown that adjunctive questions when inserted within an instructional package can enhance learning, assist in attention, and accommodate poor discriminational abilities.

It could be shown that the gifted learners are those who have acquired, developed, and internalized the mathemagenic aids that facilitate learning. But for the inattentive, it may well be, as Rothkopf has proposed that, "...It is under conditions of ineffective mathemagenic activity that treatments such as adjunct questions have produced the best results" (1974).

Traditionally, two types of adjunctive questions have been discussed: preadjunctive, meaning those questions placed in front of the material they relate to, and postadjunctive, or those that come after it.

Considering the role of the preadjunctive question, Peeck (1970) has summarized that the "experiments with pre-questions...seem to indicate that prequestions are useful when retention of certain specific information is aimed at, though depression of retention of other contents may have to be taken into the bargain."

The usefulness of post-adjunctive questions to facilitate learning and retention is well documented (Rothkopf, 1966; Rothkpf & Bisbicos, 1967; Swenson & Kulhavy, 1974). Adjunctive questions when incorporated into a printed text have assisted in the learning process.

But do they assist, as a mathemagenic aid, in audio visual lessons? And if so, where should they be placed? and what type of question is best?"

These are just the prelude to the plethora of questions that must be raised if "good design" is to be bought. Should review questions be used in a series of lessons? Should feedback be provided after a question? Should an audio visual lesson have a practice exercise? Should there be a self-evaluation at the end of the lesson and should it be different from the within program questions?

Depending upon the material being taught and the target population participating, the answer(s) may be yes or no.

The following examples are provided to illustrate this point. Recently, as part of a research study, 218 TEC lessons were observed using a checklist to monitor the presence or absence of mathemagenic aids. These lessons were representative of every TEC contractor, service school, subject matter area, and developmental year. This study indicated that there is extreme variance in the design characteristics of the lessons selected, in terms of the presence of: review questions, preadjunctive questions, postadjunctive questions, feedback, practice exercise, and self-evaluation.

SCHOOL	LENGTH OF LESSON IN MINUTES	# OF REVIEW QUESTIONS	# OF ADJUNCTIVE QUESTIONS	# OF POST ADJUNCTIVE QUESTIONS		SELF EVALUATION	PREVIOUSLY UNENCOUNTERED INSTANCE	PRACTICE EXERCISE	TIME IN MINUTES
				LOWER ORDER	HIGHER ORDER				
Infantry	30	.7	19	12	4	28%	6%	34%	5.7
Field Artillery	35.6	.3	18	12	6	92%	42%	46%	12.4
Signal	40.7	0	0	0	0	0	0	100%	25.2
Armor	31.6	1	23	15	4	74%	19%	21%	5.1
Military Police	29.9	.4	18	6	11	100%	88%	25%	8

From these few examples it can be seen that there are several different design approaches out there. Recognize that I am not proposing a singular best design strategy that combines all of the above. Rather what I am proposing, is that the appropriate mathemagenic aid(s) be inserted into the lesson at the right place as a primary design guideline to facilitate both acquisition and retention.

Certainly the purchaser of a product has to trust the professional competence of the contractor, but if you have not achieved an agreed upon design strategy that is defensible to other professionals then how can you ever buy good design? To which the rejoinder is, but all of these lessons discussed above were validated on the target population. Is validation, then, as it is presently practiced, a true indicator to the buyer that he has purchased the "best design possible in terms of efficiency and effectiveness?

As previously suggested, there is no "best design strategy." But there are suggested design strategies that should seriously be considered during the design of audiovisual lessons. If the set of objectives being taught, are distributed and sequenced over a series of lessons, then review questions to facilitate the recall of prerequisite or previously learned information are probably called for.

If there are sections of the lesson that are not intrinsically stimulating or are particularly important concepts, postadjunctive questions can perform both a backward review function and an arousal function. An appropriate combination of lower order and higher order questions may provide the best combination.

If it is important, that the student actually master the objective, and demonstrate competence, then a self-evaluation exercise at the end of the lesson can provide the opportunity to put it all together one more time, and identify any problem areas. But if this exercise is identical to the within program questions, it may test only short term memory and not acquisition of the concept.

It may well be that feedback becomes more of a crutch than an aide. Utilizing the guidelines previously identified, feedback should be included as a design element, but the unique conditions under which it may be beneficial must be considered.

Practice Exercise as a mathemagenic aid is indisputable for it allows the student to put what's in his head into his hands; and by doing it with his hands, he may well be able to solidify the actions into the process and insure retention. But practice exercise for practice exercise sake doesn't accomplish anything.

The essential ingredient emerging from all of this discussion may well be the synergistic effect of all of these design considerations when artfully employed together.

Good design is not an accident, either in terms of the buyer or the seller. It must be planned for and incorporated as a design specification within the contract. If the goal of the lesson is to achieve a 70% retention after 10 days, then it seems obvious that the validation should reflect this performance description.

The specifics of my research have been intentionally glossed over, so as to show the big picture and provide a forum for discussion. Recognizing the role and mission of the attendees at this conference, we represent a tremendously large organization that literally spends hundreds of millions of dollars per year on training. I think that all too often we fail to ask of ourselves an important question, Are we presently getting the best designed training for the money that is spent?

CONTENT VALIDATION OF CLASS A SCHOOL CURRICULA IN THE COAST GUARD

Michael J. Bosshardt, David A. Bownas
Personnel Decisions Research Institute

and

Richard S. Lanterman
U. S. Coast Guard

20th Annual Conference of the Military Testing Association
2 November 1978

Background

Coast Guard Class "A" schools train first-tour enlisted personnel for Coast Guard jobs at the E-4 level. Course work in the three schools included in this study (Aviation Electronics Technician (AT), Damage Controlman (DC), and Radioman (RM)) ranges from 15 to 28 weeks of instruction. School curricula are designed to train the rating-specific knowledges and skills outlined in the Enlisted Qualifications Manual, which specifies minimum qualification requirements for each Coast Guard specialty.

The purpose of this study was to develop and demonstrate procedures for evaluating the content validity of the "A" school curricula for preparing and selecting personnel for the three job specialties. Validation is the process of demonstrating that job selection or assignment procedures are related to job performance. When selection procedures are intended to be a representative sample of the job performance domain (the set of all tasks to be performed on the job) content validation is the most appropriate validation strategy. The recently issued Uniform Guidelines on Employee Selection Procedures state, "Where a measure of success in a training program is used as a selection procedure and the content of a training program is justified on the basis of content validity, the use should be justified on the relationship between the content of the training program and the content of the job."

The method used to demonstrate this relationship should have two properties. First, it should be as specific and precise as possible. This will increase the reliability and replicability of the results, as well as making more obvious the grounds for concluding that the fit between training and the job is either good or poor. Second, the procedure for evaluating the validity of the training curriculum should be independent of the training process. Ideally, this means that people responsible for developing a curriculum should not be asked to evaluate how well that program fits the job--there is a potential conflict of interest which could operate, however unintentionally, to influence the results. Where training personnel must be used to assess curriculum validity, the validity evaluation procedures should be made as standardized and as explicit as possible, so that the evaluation task will be as objective and clear-cut as possible for all participants.

A poor fit between training content and job requirements has several implications for personnel management. If performed tasks are not being trained, operational units perform inefficiently, and personnel time on the job will have to be devoted to task learning. If schools are training tasks which are not performed, training resources are being wasted, and some students may fail in the course because of an inability to learn tasks irrelevant to performing the job.

Procedures

The Coast Guard provided us with task lists defining the job activities for each of the three specialties. To evaluate the fit between training content and job activities, we collected three types of data.

First, we asked several instructors from each school to indicate whether each task was trained in the "A" school, and if so, how directly it was emphasized in the curriculum. To make this task more manageable, we first divided the school curriculum outline into approximately 50 homogeneous topics. Each "A" school instructor was asked to indicate how much emphasis was given to each task in those curriculum topics which he personally taught. The rating scale used for this task is shown in Figure 1. This scale has four basic levels of training: a zero or one rating indicates that a task is not trained; a two or three rating indicates that some information is presented in training that may be tangentially related to task performance, but task proficiency is not directly addressed; a four or five rating indicates that the training is directed toward task performance, but that performance is not completely developed; and a six or seven rating indicates that the training directly and specifically develops task proficiency. Two rating values were provided within each of these four basic levels to allow raters to reflect minor differences in task emphasis within level. In our curriculum evaluation analysis, we judged that a task receiving a curriculum rating of higher than three was "trained", and that tasks receiving ratings of three or lower were "not trained".

The second set of ratings we obtained was also aimed at identifying which tasks are trained. In this rating, we asked ten graduating students from each "A" school to indicate whether they could perform each of the tasks defining their specialty. Tasks they could perform were considered "learned", although some of these tasks may have been learned prior to "A" school.

The third data set we collected included ratings of whether each task was performed on the job, and if so, the relative time spent performing the task, task difficulty, and task criticality. We chose these rating factors because we felt that strong curriculum emphasis on a task could be justified if the task were time consuming, or critical, or difficult. Our own experience with earlier task inventories suggested that time spent and criticality were sufficiently independent to warrant measuring the two factors separately. Task difficulty has usually been assessed in the past by estimating the amount of training required for task proficiency, but since in this case we were evaluating whether the amount of training was indeed appropriate, we sought an independent estimate of task difficulty, without referring to training requirements. Ratings on all three factors were obtained

Instructions for Curriculum Element Contribution Ratings

Attached is a list of about 50 curriculum elements or course topics covered in your A school. First, review the list and identify those curriculum elements which you are currently teaching. In the upper right hand corner of the attached rating form are spaces for the numbers of up to 17 curriculum elements. Write the numbers of the curriculum elements you teach in these spaces. It may be helpful if you write the name of each curriculum element in the area above its number as well.

Now, consider only those tasks which you did not check in your first rating, and only those curriculum elements which you are currently training. We want to know how much each of your curriculum elements contributes to developing proficiency in each task. Use the following scale to estimate the contribution to task proficiency made by each curriculum element:

0	Unrelated
1	
2	Contributes little or only indirectly
3	
4	Makes a direct contribution or is a prerequisite to task proficiency
5	
6	Directly develops nearly complete task proficiency
7	

FIGURE 1. Rating scale for making task training emphasis ratings.

from senior enlisted personnel who supervised E-4s in the three specialties. The raters were asked to consider the activities performed by the E-4s under them, and to rate each task's time spent, criticality, and difficulty using 9-point relative scales ranging from "much below average" to "much above average". Since training emphasis could be justified for time consuming *or* critical *or* difficult tasks, each task was assigned a value equal to the highest mean rating for any of the three rating factors, and tasks with mean rating values greater than 3.0 ("below average") were considered "performed".

These data allowed us to perform three curriculum content validation analyses. First, by correlating mean ratings of training emphasis for each task with mean time spent, difficulty or criticality ratings, we were able to evaluate the extent to which the training emphasis profile matched the job rating profile across all tasks.

Second, by referring to simple ratings of whether tasks were trained in the "A" schools, whether "A" school graduates could perform tasks, and whether E-4s in the field were required to perform tasks, we were able to draw several conclusions about curriculum quality. If tasks were trained but not performed, they were considered "over trained". If tasks were rated as being performed on the job, *and* as being trained, but recent "A" school graduates indicated they could not perform them, they were considered "not learned". Finally, if tasks were performed on the job, but were *not* trained in the school and were not already within the repertoire of graduating students, they were flagged as "not trained," and should be considered for inclusion either in the "A" school curriculum or in some other Coast Guard training program such as basic recruit training.

The third validation analysis was similar to the second, but was based on continuous ratings of training emphasis and task time spent, criticality, or difficulty instead of the dichotomous performed-not performed and trained-not ratings. Again, the analysis identifies tasks that are trained (rated above 3.0 in training emphasis) but not performed (rated below 3.0 in time spent, criticality, or difficulty).

Results

Initial data analyses showed most rating reliabilities were quite high. Interrater agreement on most factors was in the upper .80s or lower .90s, with approximately ten raters.

Correlations between dichotomous ratings of whether tasks are trained and whether tasks are performed are acceptably high as shown in Table 1. For the dichotomous ratings (trained-not

Table 1

Correlations Between Task Training Emphasis and Task Job Requirements for Three Coast Guard Jobs

Job	Number Tasks	Training-Job Correlation for Dichotomous Data	Training-Job Correlation for Continuous Data
RM	403	.89	.74
AT	327	.84	.65
DC	477	.94	.82

1169

trained and performed-not performed) these correlations are .89 for RM, .84 for AT, and .94 for DC "A" school curricula. For the continuous training emphasis and job task factor ratings the correlations between school and job ratings are .74 for RM, .65 for AT, and .82 for DC "A" schools. These values suggest that the schools are training those tasks regularly performed on the job, and that the most time consuming, difficult, and critical tasks receive the most emphasis in the "A" school curricula.

Figure 2 shows some of the curriculum validation results we have obtained thus far for one of the three specialties. The data shown in this figure are proportions of training raters who indicated each task was trained, the proportions of graduating students who reported they could perform the tasks, and the proportions of supervisors who indicated the tasks were performed on the job. We more or less arbitrarily decided that, for this illustration, any task rated as "trained" by ten percent or more training experts would be considered "trained", and any task rated as "performed" by more than 30 percent of the supervisory raters would be considered "performed". We felt these values would produce a "conservative" picture of training-job fit. If as few as 10 percent of instructors indicated a task was trained, and as many as 70 percent of supervisors indicated it was performed, the task would be flagged as potentially overtrained. The first two tasks in Figure 2 are examples of a good fit between job requirements and training content: where job demand is high, training emphasis is high, and where the job demand is low, tasks are not trained. (The 60 percent of trainees who report they are able to prepare shipyard overhaul requests in task 2 have been exposed to the required forms in their careers prior to "A" school.) Of the 477 tasks in this specialty, 452 showed this kind of fit. The remaining 25 tasks in Figure 2 are those that were flagged for consideration by "A" school personnel.

Tasks that are trained but not performed are flagged in the fourth column as "over trained" (e.g., tasks 12, 16, and 32). "A" school personnel must consider these tasks and decide whether they actually are being trained, and if so, why they are trained given their low contribution to the job. Some tasks that were flagged as over trained appear to be errors by training raters more than curriculum faults. Thus, in task 32, since students learn to read diagrams and blueprints in this school, the training raters felt they were contributing to the students' ability to teach blueprint and diagram reading, and indicated that the task was trained, when, in fact, the curriculum in question does not include sessions on how to teach others to read blueprints. Other flagged tasks seem to reflect the "A" school setting. Thus, if students stood watches during their "A" school assignment, the instructors indicated that these watchstanding activities were taught at the school (e.g., tasks 421, 423, and 424). These

U.S. COAST GUARD CURRICULUM VALIDATION

DC TASK TRAINING/JOB COMPARISON
TRG PROP CUTOFF = .10   JOB PROP CUTOFF = .30

TASK		N=3 PROP TRND	N=10 PROP TRNEE ABLE	N=10 PROP JOB PFMG	OVER TRAINED	NOT LEARNED	NOT TRAINED
4	Assign DC personnel to daily tasks	1.00	1.00	.90			
7	Prepare ship-yard overhaul requests	.00	.60	.10			
12	Plan NBC drills	1.00	1.00	.30	***		
16	Prepare watch, quarters, and station bills	1.00	.90	.10	***		
32	Teach reading and drafting blueprints	1.00	.40	.10	***		
212	Operate MIG welding equipment	.00	.20	.70			***
213	Operate TIG welding equipment	.00	.20	.60			***
265	Repair portable pumps	1.00	.00	.80		***	
266	Repair furniture	1.00	.00	1.00		***	
267	Repair ladders and gangways	.33	.00	.90		***	
268	Repair ceramic tile	.00	.00	.80			***
269	Repair tile deck covering	1.00	.00	.90		***	
270	Repair sinks	1.00	.00	.80		***	
271	Repair flushing units	1.00	.00	1.00		***	
272	Repair firemain system	.00	.00	.90			***
273	Repair pressurized air system	.00	.00	.80			***
274	Repair fresh-water system	.00	.00	.80			***
275	Repair fixed CO_2 system	.00	.00	.80			***
276	Repair sanitary system	.00	.00	.90			***
277	Repair piers, camels, floats, ramps, etc.	.00	.00	.60			***
278	Repair lock and key systems	1.00	.00	.70		***	
279	Repair roofing	1.00	.00	.70		***	
280	Repair minor drainage problems	1.00	.00	1.00		***	
421	Perform duty as wheel watch	1.00	.70	.20	***		
423	Perform duty as loran watch	1.00	.40	.20	***		
424	Perform duty as teletype watch	1.00	.50	.10	***		
459	Work as diver	.00	.20	.40			***

FIGURE 2. Tasks flagged in Damage Controlman curriculum content validation analysis.

kinds of "over trained" tasks do not represent any serious problems in "A" school content.

Tasks that are performed on the job, and that are rated as being trained in the school, but which cannot be performed by graduating students are flagged as "not learned" in column five. For example, all job raters agree that task 266, repairing furniture, is performed on the job, and all "A" school raters agree that the task is trained, but none of the ten graduating students we sampled felt he could repair furniture. In many cases, such tasks can be attributed to differences of interpretation; for example, if instructors teach general woodworking skills and mention that these principles apply to furniture repair as well as small boat repair, they may feel they have "taught" furniture repairing, while students, who can't *recall* being shown how to reupholster couches, interpret the same task more narrowly. In any case, we will ask "A" school people to review these "not learned" tasks, and to decide whether they imply any weaknesses in the curriculum.

Finally, tasks that are required on the job, but which are not trained at the "A" school, and which are not already within the repertoire of graduating students are flagged as "not trained" in column six. These represent tasks which are currently being learned on the job or in later schools. The Coast Guard will review them to determine whether they should be formally trained in "A" school, in basic training, or in some other training operation, to ensure that personnel are being sent to their field assignments with all the skills they will require on the job.

Summary

In review, we have attempted to develop a procedure which the Coast Guard can use to evaluate the content validity of its Class A school curricula. The method provides an overall quantitative index of the fit between job task requirements and training task emphasis, and pinpoints specific areas of potential curriculum improvement. These specific problems will be used to stimulate discussions with "A" school personnel, not to assail them or to condemn their programs.

The results for the three specialties evaluated in this demonstration project suggest that Coast Guard "A" schools are doing an excellent job of matching their course content to critical job requirements.

EXPERIMENTAL EVALUATION
OF A
HIGH TECHNOLOGY TRAINING PROGRAM

- Arthur Kahn
Systems Development Division
Westinghouse Electric Corporation

The manufacturing of certain high technology products require that the individuals who perform the various operations should be highly skilled people. In order to acquire these skills they are usually exposed to an extensive training program. However, in performing their part in the manufacturing process, they must do more than manipulate equipment and perform highly skilled manual tasks. They must read drawings, interpret these drawings, make a record of their activity on prescribed forms at the appropriate times during the entire operation and evaluate their own performance. Usually at the completion of the training program each individual receives a certificate indicating that he or she has satisfied the performance requirements of production line and he or she is qualified to be a production worker.

At the present time in the Multi hybrid assembly area, the method for certifying that operators are qualified to work on the production line after a period of training is the following: After a given period of training the instructor starts the student doing production work of a relatively simple kind. After the student has performed this task for some time, the instructor examines this effort. If it is satisfactory, it is submitted to a Q. C. engineer for inspection. This engineer selects random samples of the effort for inspection. If they satisfy his criteria, the individual who has produced then is certified. He or she receives a card indicating that he or she has been certified. This individual is now a full-fledged production employee and is expected to produce quality work in a timely fashion.

Over the course of time it had appeared that although these individuals were certified, the data suggested that the individual performance had deteriorated from the performance at the conclusion of the training period. Although the amount of quantitative information for each worker showing the deterioration was rather tenuous, the fact that the workers had satisfied the Q. C. criteria at certification indicated that the operator could perform the highly skilled tasks satisfactorily. Thus, the question became rather obvious: Why did the performance deteriorate? The deterioration could be caused by several factors. The first was poor attitude on the part of the workers. The second was ineffective training even though they were passing certification. The third was inappropriate certification procedures. The fourth was poor job performance aids such as drawings. In the period prior to the formulation of this study, the job performance aids had been improved by various procedures. The certification procedures were those outlined in the quality control documents. These documents specify the quality of performance required so that it would appear that the certification standards are adequate. Observation of the individuals at their work station and discussion with supervisors indicate that the attitudes of the individuals are not a problem. The individuals appear to be paying attention to their work. The analysis therefore suggested that the source of the difficulty could either be in the training program or the individuals selected for the training program.

At first glance it appeared that a test should be designed to select those individuals who would benefit from training. However, such a test would involve the difficult task of validation. The task would be a difficult one because of the perceptual nature of the task that the individual had to perform. A second question, ancillary to the first was: Could a test be devised by which it might be possible to weed out individuals who could not profit from further training? In both of these questions there was the basic notion that if you intended to use the test as a selection device, there was the problem of proving the validity of the test. This last item concerned is a rather costly and unnecessary task under the present circumstance. It might be cheaper to train all the candidates and eliminate those who could not master the task. Available information suggest that individuals who find wire bonding and chip mounting incompatible with their own desires and own evaluation of their capabilities usually select themselves out i.e., they usually drop out of the training program of their own accord. Therefore, there is no need for a selection test. A third question was could the test be used to determine what and when individuals needed retaining. After analysis it became obvious that these questions could only be answered by empirical data. However, it was evident that the problem of validity should not be considered. Thus, the problem resolved itself into an evaluation of the training program.

This report is the presentation of the work that has been performed to answer these questions. An experiment had been conducted. Its aim was to evaluate the effectiveness of the training program and to determine if the device used for evaluating the effectiveness of the program could be used as a measuring tool to determine whether individuals had learned all they need to learn prior to having their work submitted for certification.

This report will cover the procedures used for preparing the subjects, description of subjects, a description of the method used to assure anonymity of subjects, a description of the substrate, the procedure for conducting the performance test, and the scoring procedure. These results of the study and

1175

the conclusions that have been derived are discussed.

In order for the program to achieve its aim it was essential that the cooperation of all the subjects should be enlisted. Therefore, an orientation meeting was held with all of the subjects, the workers in the Multihybrid assembly area. At this meeting the following material was read to them:

"This program has been prepared to evaluate a scheme for determining whether an individual has obtained from the training program all that the individual has been expected to obtain. In this program, everyone will be given a substrate that has a serial number on it. Since we will not be having everyone working on this substrate at the same time, we would appreciate your not discussing the substrate or the work with your co-workers. We should like to have you work to the level A quality rules. We want you to complete the paperwork as required but it will not be necessary to put your own initials. In fact, it would be better if you made up initials. The important part is that the initials appear in the correct number of places. No one will be able to identify the name of an individual with the performance on any substrate. This will be accomplished in the following manner:

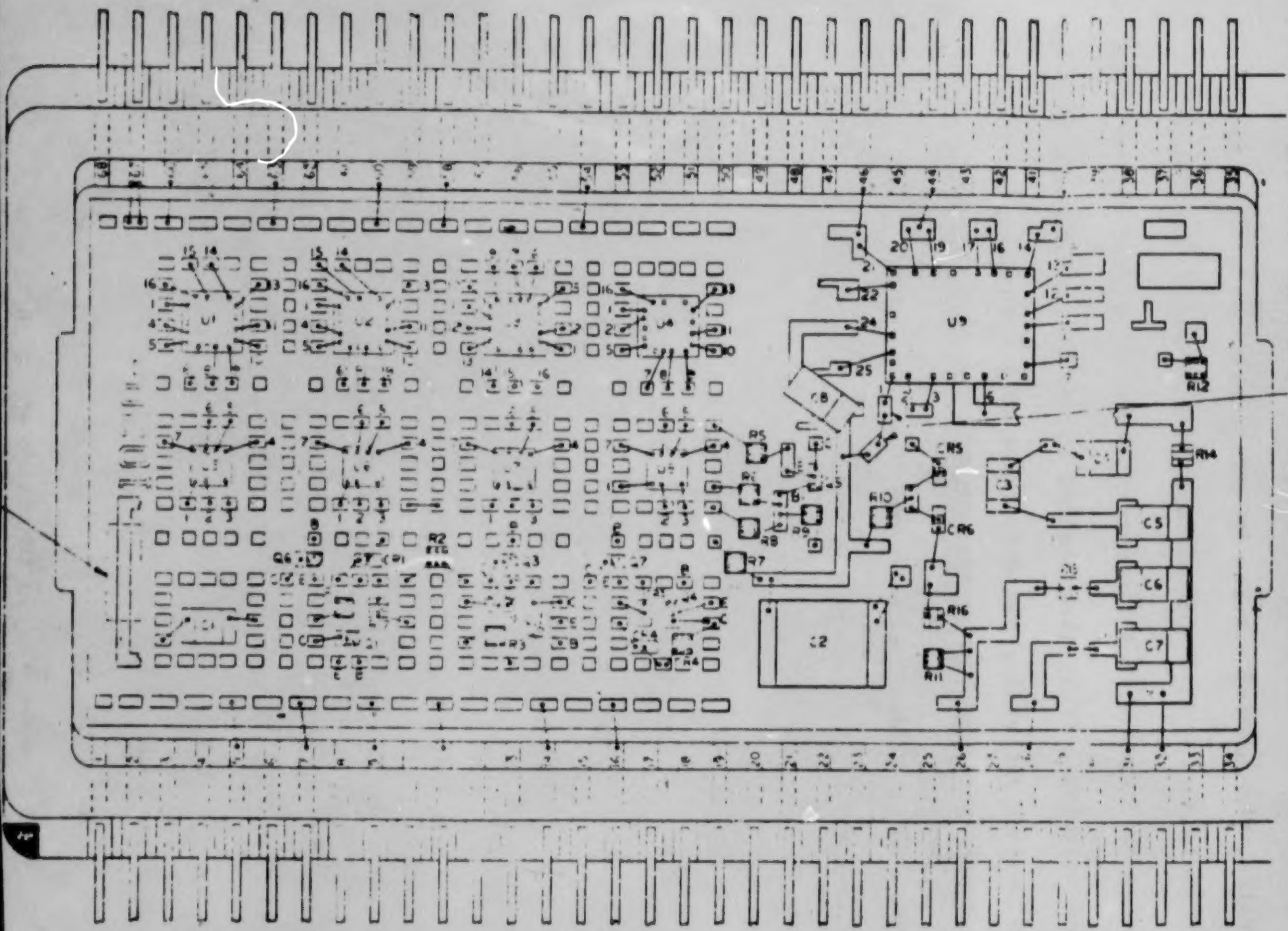
"Each name has a number assigned to it, and I have the list. We will select the first eight who will do the work shortly, by drawing the eight names out of a hat. Each slip will have a name and number on it. The individual who is selected will then put this number in the upper right hand corner of the control tag."

"After you complete the work that is required, the substrate will be checked by people selected by the Q. C. engineer. You will give the completed substrate to Mrs. Chavis (the instructor) Mrs. Chavis will give it to the engineer and he will give it to an inspector. The inspector will complete a

Figure 1. Chip Mounting

Figure 2. Wire Bonding - Thermal Compression

Figure 3. Wire Bonding - Ultrasonic



U1
U2
U4
U9
R2
R3
R4
R5
R7
R8
R9
R10
R11
R12
R14
R16
CR5
CR6
C2
C5
C6
C7
Q6

Figure 4. Substrate Drawing

check sheet on which he or she will indicate what he or she found acceptable or not. These sheets will then be given to me. I will assign the score values and calculate total scores as a percentage of the total."

"After the experiment is completed, I will tell all of you who are interested, the general results in numerical terms without talking about any single individual. If the experiment is successful, we intend to use the substrate to determine when a trainee is ready for certification or in the need of more training. It is also planned to use it to determine when individuals are in need of retraining and what phase of the training is involved."

"If you have any questions while working on the substrate, please ask Mrs. Chavis or Chuck Luedtke,(the designer of the experimental substrate) Please work as quickly as you can but remember the emphasis should be on quality."

"Do you have any questions?"

The subjects were experienced wire bonder and chip mounters who were labor grade 7's. They were all female. All except two were experienced in performing both wire bonders and chip mounters. One of these was a chip mounter and the other, a wire bonder. Although all were experienced personnel at the time the study had begun, some were performing other tasks required of labor grade 7 other than wire bonding and chip mounting, e.g., 100X inspection. Figure 1, 2 and 3 show individuals performing the chip mounting and wire bonding tasks. The individuals were selected at random, until all had completed the test substrate. The only limiting factor to this procedure was the number of individuals who were selected at the time. For example in the first group, there were eight people on the first shift and two on the second shift; in the second batch there were four in the first shift and two on the second shift.

As the remaining numbers reduced, the number was reduced so that at the end of the study there were individuals working by themselves.

All subjects were assigned a number, the only individuals who knew which numbers were assigned to each individual were the individual; Mrs. Chavis, who assigned the numbers, and the experimenter. The only individual who was able to associate a score with an individual was the experimenter.

The control card that the inspector examined contain only a number. Since the inspector received the substrate and paperwork from the instructor. It was not possible for the inspector to associate any substrate with any number or name. Once Mrs. Chavis gave the substrate to the inspector, she did not examine the score sheet but submitted it to the engineer who monitored the test to assure that the tantalum capacitor (a particular component) had been properly mounted. The inspector's recording sheet contained only the serial designation of the substrate.

The substrate used for this study was a special designed unit. It contained resistors, capacitors, diodes, IC chips and tantalum capacitors. It was designed so that the individuals would be required to exercise judgement to determine the sequences of chip placement as well as the order of placing the wire bonds. The chips were of different sizes. In addition to the units to be assembled there were notes on the drawing that had to be followed. Figure 4 is a drawing of the basic substrate. Figure 5 is a photograph of the completely assembled device. Figure 6 is an enlarged photograph of a section of the completed device to show the kinds of connections that had to be made.

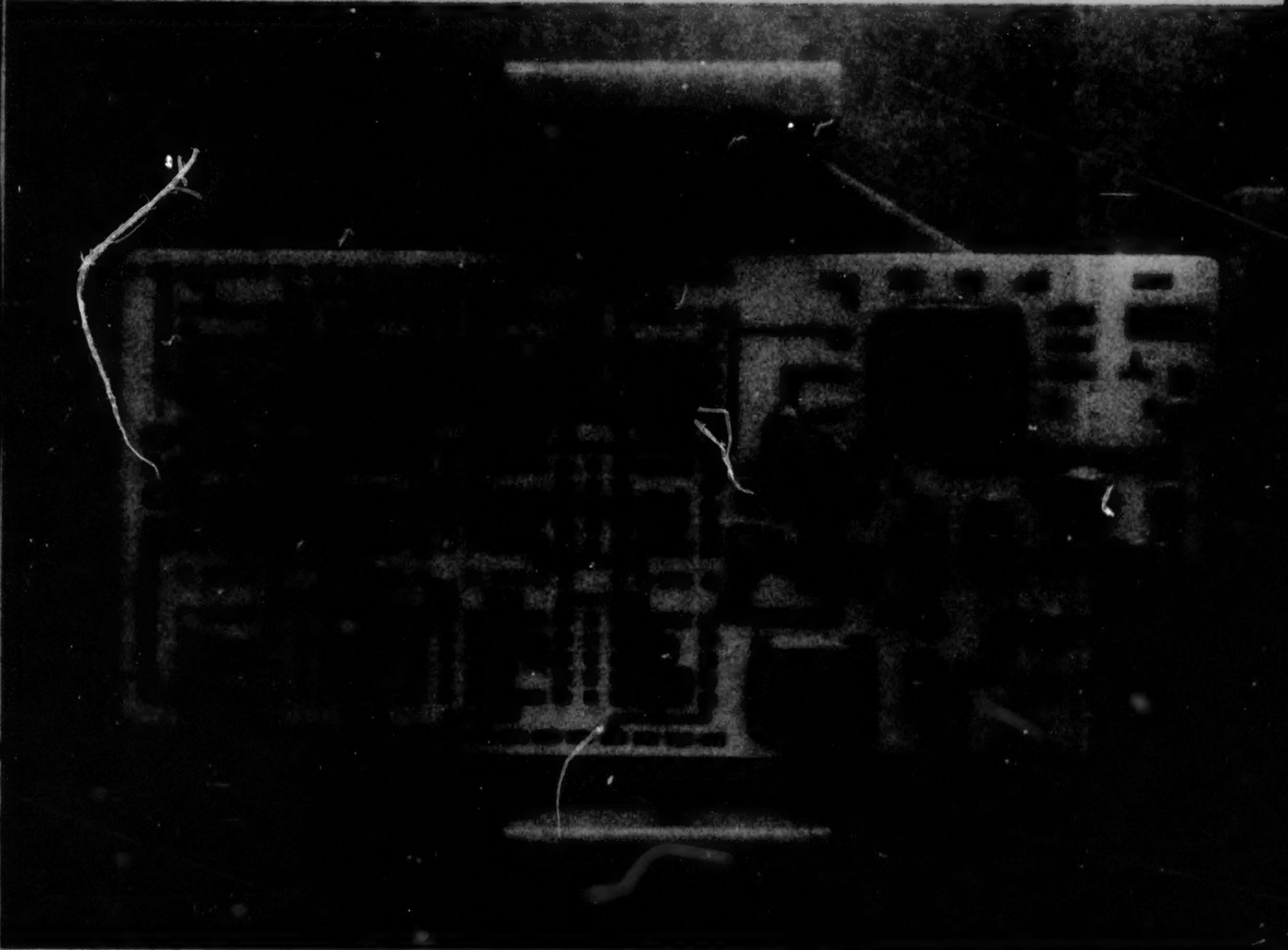


Figure 5. Completed Substrate

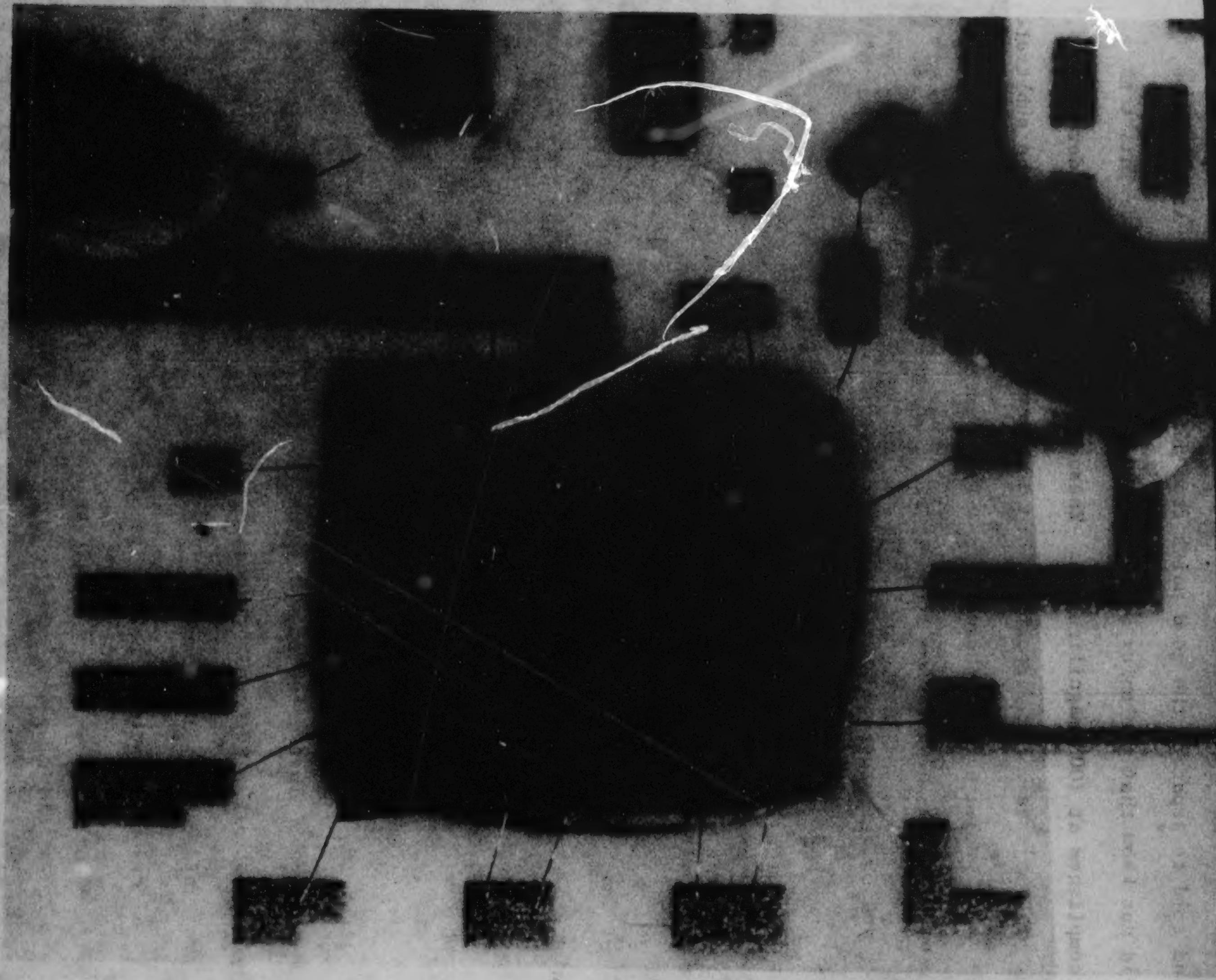


Figure 6. Enlarged Section of Substrate

Prior to the beginning of the experiment, a random selection of the individuals who were to be subjects in the experiment were given a number of parts and told to separate the good units from defective units. At this time, they did not know that the units would be used in the experiment. This task was accomplished at 100X magnification. Those parts that were considered good by this process became the parts that were used in the assembly of the substrate.

The subjects were provided with (1) a substrate, (2) a drawing showing how the [illegible] was to be assembled (fig. 4) and (3) a package containing parts selected following the procedures previously described. They also received a data package that contained the necessary paperwork that normally accompanies a job. This package contained a serialized inspection control tag, a continuation control tag, a serialized allowable rework tag, and a serialized part traceability tag. After receiving these materials, the individual proceeded to chip mount and wire bond as required. They were instructed to work until they completed the task. If they had any questions of any kind, they were instructed to bring the questions to the instructor . She would either answer the question or provide whatever was required. The subjects were also instructed to bring the completed substrate to the instructor. Since the process required "curing" (being placed in a furnace for a period of time) the substrate at the appropriate time, they were instructed to bring the substrate to her at the appropriate time. The substrate were then returned to the individual so that the tantalum capacitors could be mounted after the "curing."

After the individual had completed the mounting of the tantalum capacitors she gave the completed substrate and associated paperwork to the instructor who then gave the material to the inspector who had been assigned to the program. This individual then evaluated the substrate in accordance with level A Q. C. criteria, and recorded the result on the evaluation sheet that had been prepared. Figure 7 is the evaluation sheet for the chip mounting task. Figure 8 is the evaluation sheet for the wire bonding part of the task. The inspector wrote the word "yes" in the blank if the criteria had been satisfied and "no" if they had not been satisfied. After the inspector had evaluated the substrate and had recorded the results the substrate and paperwork was returned to Mrs. Chavis. She gave the substrate and paperwork to the engineer who had the resistance of the tantalum capacitor mounting measured by the test section personnel. A resistance of less than .2 ohms was required. The result of this test was recorded on the evaluation sheet. The purposes of this test was to determine whether the capacitors had been properly mounted. The completed evaluation sheets, paperwork and substrate were then delivered to the scorer who proceeded to provide a numerical value to the evaluation that had been accomplished.

At the time of development of the evaluation sheets for both the chip mounting, each item on the evaluation sheet was given a point value by the manufacturing engineer. The assigned value depended upon his judgement of the importance of the task being evaluated. These points were recorded on a separate sheet. They did not appear on the sheet on which the inspector recorded his results. After each item on the sheet had been scored according to the corresponding point value, the total number of points was obtained for chip mounting and wire bonding. If the individual performed all tasks correctly on the chip mounting task, he or she obtained 132.5 points and 134.5 points, on the wire bonding task. The final score obtained in each case was the actual number of points obtained divided by the number possible expressed as a percentage.

CHIP MOUNTING SHEET

Serial No. of Substrate ____________________

Eutectic Mounting Component	Orientation to 7AB assy dwg; 75% eutectic flow around chip, enough room to wire bond and to dep.	No* Damage	(Includes no chip outs, damage metal eutectic material on top of chip)
Q1	______	______	
Q2	______	______	
Q3	______	______	
Q4	______	______	
Q5	______	______	
Q6	______	______	
Q7	______	______	
CR1	______	______	
CR2	______	______	
CR3	______	______	
CR4	______	______	
CR5	______	______	
CR6	______	______	

Adhesive Mounting Component	Epoxy Visible around 75% of Chip	Correct Epoxy Used	Mech. Damage	Excess Epoxy	Wrong Orientation, Location or Mission
U1	______	______	______	______	______
U2	______	______	______	______	______
U3	______	______	______	______	______
U4	______	______	______	______	______
U5	______	______	______	______	______
U6	______	______	______	______	______
U7	______	______	______	______	______
U8	______	______	______	______	______
U9	______	______	______	______	______
CR1 thru 3	______	______	______	______	______
CR4 thru 6	______	______		______	______
C2	______	______		______	______
C1, C3, C4	______	______		______	______
Q1 thru 3	______	______	______	______	
Q4 thru 7	______	______	______	______	______
R1 thru 6	______	______		______	______
R7 thru 12	______	______		______	______
R14 thru 17	______	______		______	______
TABS					

Conductive Adhesive Components	No shorts by conductive epoxy		
U1	______	U4	______
U2	______	U5	______
U3	______	U9	______

A All lot entries made ______

B U8 mounted per RN ______

C Dielectric paste over track ______

D Q7 and Q6 mounted to track shorting to track ______ (test to verify)

E No mechanical damage to subs. ______

F Paperwork correct ______

	Joint Resistance less than 2 ohms	Satisfy Visual Criteria	No Epoxy Shorting	No Damage to Assembly
C5	______	______	______	______
C6	______	______	______	______
C7	______	______	______	______
C8	______	______	______	______

Indicate acceptance by:
Y - YES
N - NO

Figure 7. Scoring Sheet

WIRE BONDING SHEET

Serial No. of Substrate ________________
Correct from Drawing

Component	Score Bonded Correctly*	General	
U8	____	Wired so tantulums can be mounted	____
U9	____	Wire loops uniform appearance	____
U1 to U7 (each)	____	R1 and R4 not bonded	____
CR5 & 6	____	Wires not mashed	____
CR3 & 4	____	R11 bonded from a center tap	____
CR1 & 2	____	No mechanical damage to chips or substrate	____
Q3 & 4	____		
Q5 & 6	____	Penalties	____
Q7	____	One pigtail or spur not removed	____
Q1 & 2	____	Two pigtails or spurs not removed	____
R1 to 6	____	Three pigtails or spurs not removed	____
R7 to 12	____	Four or more pigtails or spurs not removed	____
R14 to 17	____		
C1	____		
C2	____		
C3	____		
C4	____		
Jumper 1	____		
Jumper 2	____		
TABS	____		

Ball Bond Pla	More than 75% of ball on pad	Ball infilet	Ball bond short	Sliding Bonds
U1	____	____	____	____
U2	____	____	____	____
U3	____	____	____	____
U4	____	____	____	____
U5	____	____	____	____
U6	____	____	____	____
U7	____	____	____	____
U8	____	____	____	____
U9	____	____	____	____
Q1	____	--	____	____
Q2	____	--	____	____
Q3	____		____	____
Q4	____	--	____	____
Q5	____	--	____	____
Q6	____	--	____	____
Q7	____	--	____	____
R14 to 17	____	--	--	--
R1 to 12	____	____	--	--
C1, C3 & C4	____	--	--	--
C2	____	--		--
CR1 & 2	____	--	--	--
CR3 & 4	____	--	--	--
CR5 & 6	____	--	--	--

Bonded Correctly*
- Wire Placement
- Ball Configuration
- Stitch Configuration
- Wire Length
- Wire Damage

Indicate acceptance by:

Y - YES
N - NO

Figure 8. Scoring Sheet

1158

The results of the study are the percentage of total points that were obtained by the individuals in the wire bonding and chip mounting tasks. Table I shows the mean and standard deviation for the different tasks and the average of both tasks.

Table I

Average Performance Measures and Their Variability
(Percent of Total Points)

	Mean	S.D.
Chip Mounting	87.33	5.46
Wire Bonding	85.33	7.19
Average	86.33	5.53

The mediam score was 87.5%.

Each score sheet was examined to see if a pattern of errors could be extracted from the results. The analysis of the wire bonding task showed that 88% of the individuals did not bond R11 correctly. 42% did not have wire loops with uniform appearance; 33% did not remove one pigtail; 37% were criticized for having mashed wires. These were Q. C. criteria that they had to keep in mind. A similar analysis was made of the chip mounting task. There were two different mounting methods required. The error rate of Eutectic mounting (an alloy forming technique) was approximately 20% while the error rate of adhesive mounting was 1%. 88% of the people did not complete the paperwork correctly. This value represented 21 of the 24 individuals. 96% did not mount U6 correctly as required by the RN. (Engineering Change) Almost 50% of the people did not use the correct epoxy in mounting U6.

Although the inspector found many instances of mechanical damage the error rate was relatively small.

The average score of 86+ indicates that as a group, all of the individuals performed very well. When consideration is given to the fact that a majority of individuals lost points because the paperwork was not correct and/or they did not comply with the RN, the average would have been greater than 90. This analysis suggests that the individuals were capable of performing the chip mounting and wire bonding tasks very well. The major problems other than the RN in the wire bonding task is related to meeting Q. C. criteria. 42% did not meet the uniform appearance criteria and 37% were criticized for having mashed problems and mechanical damage. Wires can get mashed during the movement of the substrate and/or the movement of the capillary with the wire. At the same time it is possible for chips to be damaged. One of the inspector's tasks was to discover mechanical damage. An examination of the inspector's finding and the substrates indicate that it is possible that some of the points lost because of mechanical damage were due to defective parts in the kits even though all bad units were supposed to have been removed. These findings suggest that the individuals who made the determination were not sufficiently aware of the criteria for rejecting defective components. That this state of affairs should exist is not unexpected since these individuals have not had any organized training in the recognition of defective units at 100X. It would therefore appear that training in the recognition of defective material should become part of the training of these individuals.

The results suggest that more attention during training should be given to the correct accomplishment of the necessary paperwork, the careful reading of drawings, calling attention to the information in the notes and the assuring that the proper drawings and/or revision notices are available. It is the responsibility of the individual worker to determine whether the paperwork indicates that the proper drawings are at hand. In this experiment, very few individuals recognized that an RN was required. It has been argued that the individuals had been instructed not to ask questions. They were in fact told not to ask their associates but they were encouraged to ask Mrs. Chavis if something were amiss.

In any event, the importance of training effort on paperwork became evident in the course of an extraneous exercise that occurred independent of the experiment. Mrs. Chavis was instructing an individual while the experiment was being conducted. During this period, the individual completed the instruction and completed the certification process satisfactorily. Mrs. Chavis then gave this individual the test substrate as an experiment as an innovative part of her training program. At the completion of the substrate she asked the individual what she had learned. The individual responded with the following remarks: (1) "Arranging and locating the chip on the plain substrate without metallization. (2) "Using your footnotes and flags." (3) "Making your revision." (4) "Understanding your paperwork." (5) "Listing lot numbers." (6) "Checking orientation when mounting transistors." This ad hoc experiment suggesting that the finding of deficient paperwork in the results is not an artifact is but a true representation of the performance of the individuals. It suggests that either the individuals are not aware of what is correct or they have been inattentive to this aspect of the task. Since so many individuals made paperwork errors it does not seem reasonable to attribute the errors to inattention.

This extraneous "experiment" also shows that the beginning chip mounter needs training in arranging chips when the substrate does not provide a metallized tab on which to mount it.

Earlier paragraphs have indicated that individuals have had little formal training on the recognition of defective parts at 100X. Similarly, they receive no training in the general examination at 100X. It would seem unreasonable to have individuals perform work at 30X then have it examined at 100X and then criticize the individuals for not performing good work when they are unable to tell good from bad at 30X because the defect only becomes visible at 100X.

The results also show that there is little difference in performance between chip mounting and wire bonding, the basic task of these individuals. The difference slightly less than two two percentage points. However, the error rate for Eutectic mounting was about 20 times that of the adhesive mounting. This difference could have been due to the fact that the chips used were training chips. However, similar chips were used in both processes. It is conceivable that the nature of the Eutectic mounting is such that the judgements the individuals have to make and the variability that could occur in the equipment as contrasted to that which could occur in adhesive mounting could account for this difference. The data show that 96% did not use the correct epoxy to mount U6. This error was a result of not requesting the RN.

The analysis thus far suggests that the training program should be changed in three fundamental ways. The first fundamental change should be more emphasis on correct examination of the paperwork and emphasis on the correct manner of completing all the paperwork correctly. There is no reason why the individual who can perform the complex task of wire bonding and chip mounting should make mistakes on paperwork. The individuals during training should be given practice completing the paperwork during separate exercises and then this kind of practice should be integrated into their work completing substrates. The second fundamental way the training should be changed is that additional training substrates should be developed and used. At least two should be developed, one more difficult than the other. These should include the use of RN's and the correct completion of paperwork. The third way is that the individuals should be given more practice in discriminating good from bad work. The final significant item of the results was the fact that 47% of the individuals failed the test of the tantalum capacitors. Although the units passed the visual inspection, they did not pass the electrical test. This test demonstrated that the mounting of the capacitor was sufficiently complete so that there was little electrical resistance in the bonding materials. This result suggests that a way must be found to allow the operators to be able to evaluate the quality of the eledtrical connection by the amount and distribution of epoxy.

From the learning theory point of view, the results indicate that it is not sufficient to give the individual practice in the task that must be performed but the individual must gain experience of performing this task in a different context. This means that the individual learns to vary the basic task as the function of different requirements from trial to trial. This state of affairs existed when Mrs. Chavis, the instructor, used the test substrate during her training program with new employees. It could therefore appear that the individuals need to get practice in the basic task but they must develop a large repertoire of responses before they can be successful production workers.

This particular experimental program demonstrated that a properly developed testing program will not only indicate where training programs can be improved but they provide a mechanism for applying what is known about changing human behavior.

SECTION 14

INSTRUCTIONAL EVALUATION AND TEST DEVELOPMENT

THE INSTRUCTIONAL QUALITY INVENTORY: INTRODUCTION AND OVERVIEW

John A. Ellis, Wallace H. Wulfeck II,
Navy Personnel Research and Development Center

Robert E. Richards, Norman D. Wood
The Pennsylvania State University

and
M. David Merrill
Courseware, Inc., San Diego, Ca.

INTRODUCTION

Problem

Modern military instruction is developed according to a systematic method called Instructional Systems Development (ISD). The order of development involves:

1. Job/task analysis leading to specification of training objectives;
2. Development of tests to measure student progress toward the objectives;
3. Design of new instruction and/or adaptation of existing instruction to achieve the objectives;
4. Implementation of the training program;
5. Evaluation and feedback for course maintenance.

Various military activities are using this model to develop many of their training courses. There is a need in ISD for quality control and/or evaluation procedures so that (a) quality can be maintained throughout instructional design so that errors early in development are not magnified as development proceeds, (b) existing materials can be evaluated with respect to newly derived training objectives for purposes of adaptation or revision, (c) deficiencies in performance of course graduates can be traced to possible deficiencies in instructional materials, and (d) instructional materials obtained through contract efforts can be evaluated for purposes of acceptance.

Purpose

The purpose of this research and development effort was to develop quality control/evaluation procedures, for use by military instructional design and development personnel, for the three main products of an instructional development effort, namely objectives, tests, and instructional materials or presentations. The Instructional Quality Inventory (IQI) is the result of this effort.

Background

The Instructional Strategies Diagnostic Profile (ISDP) (Merrill, *et*. *al*., 1977a) was originally developed by Courseware, Inc. under contract to the Navy Personnel Research and Development Center to provide a method for diagnosing defects in instructional materials. On the basis of empirical studies designed to validate the procedures (Merrill, *et*. *al*., 1977b; Merrill & Wood, 1977), and workshop trials of the procedures (Wood, Ellis, & Wulfeck, 1978), the ISDP was extensively revised. The revised version, entitled the Instructional Quality Inventory (Ellis, Wulfeck, *et*. *al*., 1978) has also been extensively field-tested. On the basis of these tests, the IQI was again revised to be maximally useful to instructional design and development personnel, and to be tailored to the needs of instructional design activities. This document provides an overview of the latest revision of the IQI.

Overview

The IQI procedures were designed to parallel and supplement the ISD process. The procedures are based on a system for classifying objectives, test items, and instructional presentations (the three main products of instructional development). Classification is determined according to (a) what the student is required to do with the information he learns, and (b) what type of information the student is learning. The IQI procedures include the following:

1. Since all ISD steps depend on careful specification of learning objectives, the first IQI procedure is to assure the *adequacy* of objectives. This is done by classifying each objective, and judging whether or not it accurately reflects the intended student performance after training.

2. The next step is to insure that tests accurately measure progress toward the objectives. This is done by assessing *consistency* between test items and their associated objectives. Essentially, each test item must be classified in the same way as its objective. After test items and objectives are consistent, the adequacy of the test items is assessed.

3. The next step is to insure that the instructional presentation is *consistent* with the objectives and tests; then to insure that the instructional presentation is *adequate* according to psychological principles of learning.

What the IQI is and is not

The IQI is basically a logical analysis method which makes use of sound principles of training and the psychology of learning and instruction. The IQI deals principally with cognitive and psychomotor instruction rather than attitudes, motivation, etc.

Using the analysis procedures of the IQI to rate the consistency and adequacy of an instructional program, and making revisions on the basis of these analyses, can greatly reduce the time and effort needed to validate and revise an instructional course or system. However, although the IQI can reduce the need for validation on real students, it does not entirely eliminate the need for empirical tryouts.

The IQI is a method for product evaluation, not process evaluation. Regardless of the development methodology used to produce the objectives, tests, or instruction, the IQI can be used to evaluate the quality of the products. The IQI criteria can be kept in mind during the development of instruction, but the IQI is intended as a supplement to ISD, not a replacement for it.

The IQI is intended for use by people with knowledge of ISD; it cannot be used by untrained personnel. Also, the application of the IQI depends upon a good task analysis, or the availability of subject-matter experts, and preferably both. This is because the IQI assumes that what needs to be taught has already been determined.

Organization of this paper

The following section of this paper is an introduction to the IQI procedures. It is designed to acquaint managers of instructional development efforts, evaluators of instruction, contract monitors, etc., with the IQI. While it provides a substantive overview of the IQI process, it is not a complete IQI training program.

There are three other volumes in the IQI series which will be available in early 1979. These are

1. a User's Manual, which contains a complete description of all IQI procedures, and examples of their use.

2. a Training Workbook, which contains additional examples and practice on the IQI procedures.

3. a Job Performance Aid, which contains a brief version of each IQI procedure.

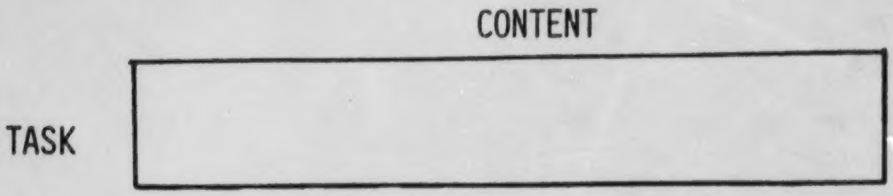
INSTRUCTIONAL QUALITY INVENTORY PROCEDURES

THE CLASSIFICATION SYSTEM:

The following classification system is used in all IQI procedures. It is applied to the three main parts of instruction: objectives, tests, and instructional presentations.

Each objective, test item, or piece of presentation, can be classified according to:

1. What the student must do, i.e., the TASK to be performed, and
2. The type of information the student must learn, i.e., the instructional CONTENT.



In the IQI, these two classification dimensions have been combined to form the TASK/CONTENT MATRIX.

THE TASK DIMENSION:

There are two main TASKS a student can perform:

1. He can REMEMBER information, or
2. He can USE the information to do something.

REMEMBER	
USE	

EXAMPLE:

Here are two test items:

1. The symbol for resistor is ___________.
2. Using your knowledge of electronic theory, what would happen in the circuit shown below if the load resistance were shorted?

These two test items differ with respect to what the student is supposed to do (TASK). *In number 1, the student has to* REMEMBER *something, and in number 2, the student has to apply or* USE *his knowledge in a new situation.*

THE CONTENT DIMENSION:

There are five types of CONTENT:

	FACT	CONCEPT	PROCEDURE	RULE	PRINCIPLE
remember					
use					

FACTS are simple associations between names, objects, symbols, locations, etc.

CONCEPTS are categories or classifications defined by certain specified characteristics.

PROCEDURES consist of ordered sequences of steps or operations performed on a single object or in a specific situation.

RULES also consist of ordered sequences of operations, but can be performed on a variety of objects or in a variety of situations.

PRINCIPLES involve explanations or predictions of why things happen in the world. That is, they concern predictions or interpretations based on theoretical or cause-effect relationships.

NOTE: Facts can only be remembered. The others can be remembered or used.

EXAMPLES:

The following examples illustrate the five content areas for the REMEMBER *task level:*

REMEMBER FACT	*1. The symbol for resistor is________.* *2. The student will list the names of the parts in the wind indicating instrument.*
REMEMBER CONCEPT	*1. List the defining characteristics of a jet pump.* *2. The student will define the various kinds of clouds (cumulus, stratus, etc.).*
REMEMBER PROCEDURE	*1. List in order the steps for cleaning an M-16 rifle.* *2. The student will describe the procedure for preparing and sending a radio message.*
REMEMBER RULE	*1. List the steps involved in finding the rhumb-line course between two points on the earth.* *2. The student will state the general rule for solving for circuit current, given voltage and resistance.*
REMEMBER PRINCIPLE	*1. State the principles of electron movement in a semiconductor junction.* *2. The student will recall the reasons why hydraulic fluid contamination must be avoided.*

Facts can only be remembered, but for the other content types, the student may be asked to USE *his knowledge to classify, perform, solve, or predict. The following are examples of the* USE *task level for all content types except facts:*

USE CONCEPT	*1. Which of the pumps aboard ship are jet pumps?* *2. Given photographs of clouds, the student will sort them according to type (cumulus, stratus, etc.).*
USE PROCEDURE	*1. Clean an M-16 rifle.* *2. The student will prepare and send a radio message.*
USE RULE	*1. Calculate the rhumb-line course from Pearl Harbor to Long Beach.* *2. Given the values for voltage and resistance, the student will calculate the current flow.*
USE PRINCIPLE	*1. Describe the theoretical movement of electrons in a PNP transistor.* *2. The student will predict what is likely to occur if the landing gear fluid were contaminated.*

THE USE LEVEL CAN BE FURTHER DIVIDED INTO TWO TYPES:

1. USE-UNAIDED in which the student has no aids except his own memory.
2. USE-AIDED in which the student has a job aid for performing the task.

For this level, the nature of the aid depends on the content type:

For USE-AIDED CONCEPTS *the aid should consist of a decision strategy, including each critical characteristic, and the decision to be made according to presence or absence of that characteristic. In simple cases, the aid may only include a list of characteristics; the decision strategy is then implied.*

For USE-AIDED PROCEDURES *the aid would be a list of steps to be performed.*

For USE-AIDED RULES *the aid would be at least a statement of the formula or rule to be applied, and could include guidelines for when and how to apply it.*

For USE-AIDED PRINCIPLES *the aid would also be at least a statement of the principle, and could include guidelines for when and how to apply it.*

EXAMPLES:

USE-AIDED:	*A pilot's preflight checklist is a* USE-AIDED *procedure. The pilot does not have to remember the steps or their order because they are on the checklist. The pilot does need to perform the steps correctly.*
USE-UNAIDED:	*"The student will field-strip an M-16 rifle." Here, the student must remember the steps in the correct order, and perform them correctly.*

In summary, the REMEMBER level involves "pure" remembering,

the USE-UNAIDED level involves remembering what is to be used, and then using it, and

the USE-AIDED level involves "pure" using.

THE ENTIRE TASK / CONTENT MATRIX IS SHOWN BELOW:

	FACT RECALL OR RECOGNIZE NAMES, PARTS, DATES, PLACES, ETC.	CONCEPT REMEMBER CHARACTERISTICS, OR CLASSIFY OBJECTS, EVENTS OR IDEAS ACCORDING TO CHARACTERISTICS	PROCEDURE SEQUENCE OF STEPS REMEMBERED OR USED IN A SINGLE SITUATION OR ON A SINGLE PIECE OF EQUIPMENT	RULE REMEMBER OR USE A SEQUENCE OF STEPS WHICH APPLY ACCROSS SITUATIONS OR ACROSS EQUIPMENTS	PRINCIPLE REMEMBER, OR INTERPRET / PREDICT, WHY OR HOW THINGS HAPPEN, OR CAUSE-EFFECT RELATIONSHIPS
REMEMBER - RECALL OR RECOGNIZE FACTS, CONCEPT DEFINITIONS, STEPS OF PROCEDURES OR RULES, STATEMENTS OF PRINCIPLE					
USE-UNAIDED - TASKS WHICH REQUIRE CLASSIFYING, PERFORMING A PROCEDURE, USING A RULE, EXPLAINING OR PREDICTING WITH NO AIDS EXCEPT MEMORY.					
USE-AIDED - SAME AS USE-UNAIDED, EXCEPT JOB AIDS ARE AVAILABLE.					

Any objective, test item, or piece of instruction will be classifiable in one and only one cell of the matrix above.

This matrix is used in all IQI steps.

1203

OBJECTIVE ADEQUACY:

The first step in the IQI procedure corresponds with the development of training objectives. The procedure described below is used to determine if each objective is adequate for further instructional development.

Objectives are ADEQUATE if they satisfy three general criteria:

1. **Is the objective CORRECTLY STATED?** *Does the objective include statements of actions the student is to perform after training, the conditions under which the performance is expected, and the standards which the performance must meet? If even one of these parts is missing, the objective is inadequate because training for it cannot be designed or evaluated.*

   EXAMPLE: *Inadequate objective: "The student will prepare a standard Navy message." This is inadequate because it does not specify either the conditions (given a typewriter? TTY?) or the standards (how fast and how many errors).*

2. **Is the objective CLASSIFIABLE on the task/content matrix?** *If the objective cannot be classified, this means that the action the student is to perform is not stated clearly enough so that we know what the student is to do. Training cannot be designed or evaluated in these circumstances.*

   EXAMPLE: *The objective "The student will learn repair procedures for the XYZ radar set" is not classifiable. It is not clear whether the student should remember the procedures or actually use them.*

3. **Is the "intent" of the objective APPROPRIATE for the purpose of the course?** *The actions, conditions, and standards specified in the objective should be as close as possible to the actions, conditions, and standards of the task to be performed on the job after training.*

   In addition, it is assumed that the ultimate "intent" of any training program is to teach the student how to do something (i.e. USE level). Therefore, there must be a USE-level objective for each REMEMBER objective. (Facts are a special case: Although facts are not used, they often must be taught to provide a knowledge base for a later use-level task. Therefore, in order to justify teaching facts, they must support some use-level objective.)

   Conversely, USE-UNAIDED tasks should be taught at the REMEMBER level before being taught at the USE level. Therefore, just as every REMEMBER objective should have a corresponding USE objective, every USE objective should have a previous REMEMBER objective.

   EXAMPLE: *The objective "The student will identify the connection of a voltmeter to measure the voltage across a component by selecting an illustration" is not appropriate for the intent of the course. The student will not see possible illustrations of connections on the job, but will be required to set up the connection, thus the action should be revised.*

1204

TEST CONSISTENCY AND ADEQUACY:

Once objectives are adequate, test items can be developed. The next IQI step is the quality control step for test development. This step involves determining whether test items are CONSISTENT *with objectives, and whether each item is* ADEQUATE.

A test item is CONSISTENT with its objective if:

1. The ACTION (TASK/CONTENT level) of the test item is the same as that of the objective.

2. The CONDITIONS under which the item is administered are as close as possible to those of the objective.

3. The STANDARDS in the test item, or the STANDARDS for scoring the item, are as close as possible to the standards in the objective.

__EXAMPLE:__ Objective: Given the necessary tools and an operator's manual, the student will set up and operate a double-acting reciprocating pump, in five minutes and according to the manual specifications.

Inconsistent test item: "List the steps of procedure for starting, operating, and stopping a double-acting reciprocating pump."

This test item is inconsistent, because its TASK/CONTENT is REMEMBER-PROCEDURE instead of USE-AIDED-PROCEDURE. Notice that the action the student is to perform in the test is not the same as the action required in the objective.

Consistent test item: "Use the operator's manual, and necessary tools to set up and operate a double-acting reciprocating pump. You will pass this test if you complete this task within 5 minutes, in accordance with the manual specifications."

This test item is consistent with the objective. Notice, however, that if either the conditions or grading standards had been left out, the item would have been inconsistent.

A test item may be consistent with its objective, but may still not be an adequate item. TEST ADEQUACY *depends on a number of criteria that items must satisfy.*

After a test item is consistent with its objective, the test item is ADEQUATE if:

1. The item is clear and unambiguous.
2. The item does not give away its own answer or the answer to any other item on the test.
3. The format of the test item is appropriate for the TASK/CONTENT level.
4. Other adequacy concerns covered in the IQI manual are met.

EXAMPLES:

1. *"Which of the following..." is ambiguous because it does not say "choose all that apply" or "choose the best...".*
2. *"The steps in the procedure for operating a jet pump are listed below. Arrange them in the correct order." This is an innapropriate format for REMEMBER-PROCEDURE because the student doesn't have to remember the procedure, only recognize it.*

Note: Recognition items (multiple choice, matching, true-false) are usually NOT *appropriate test formats for REMEMBER-level objectives. This is because these items do not reflect typical job-performance requirements.*

Multiple-choice, matching, and true-false items are appropriate for concept recognition, and can be appropriate for USE level objectives if they are carefully designed. However, for USE level objectives, "hands-on" performance tests are usually most appropriate.

PRESENTATION CONSISTENCY:

At this point in the IQI process, objectives are adequate, test items are consistent with objectives, and test items are adequate. The next instructional design step is to prepare the instructional materials or presentations. The next IQI step is to insure that the presentations are CONSISTENT *with the objectives and test items, and are* ADEQUATE.

In the previous section, determining test-objective consistency involved comparing each test item with its related objective. Determining PRESENTATION CONSISTENCY, *on the other hand, involves checking whether or not each of the INSTRUCTIONAL COMPONENTS required for a given objective-test item is present. There are different types of instructional components. In order to insure consistency, the appropriate components must be present for each TASK/CONTENT level. Not all task/content levels require all components.*

The Instructional PRESENTATION COMPONENTS are:

1. STATEMENT: *The instruction tells the student something he must learn.*
2. EXAMPLES: *The instruction shows the USE of content (concept, procedure, rule, or principle).*
3. PRACTICE: *The student practices REMEMBERING or USING the content, and is given feedback.*

1207

PRESENTATION COMPONENTS:

STATEMENT Component: The instruction presents a statement of a fact, a concept definition, the steps of a procedure or rule, or a statement of a principle.

EXAMPLES:

1. *"The characteristics of a typical jet pump include...." (concept definition).*
2. *"The procedure for changing a gasket in a check valve is ..." (procedure definition).*
3. *"To determine voltage, multiply current by resistance." (rule statement).*

EXAMPLE Component: The student is told or shown how a statement of a concept, procedure, rule, or principle applies in a specific case.

EXAMPLES:

1. *"The XXYZ pump is a double-acting reciprocating pump because it has the particular characteristics noted on the diagram below." (concept example).*
2. *"Let's see how OHM'S LAW applies in a specific case...." (rule example).*
3. *"The Navy's victory at Midway in World War II illustrates the value of cryptologic intelligence because..." (principle example).*

PRACTICE REMEMBERING Component: The student is asked to supply part or all of a fact statement, concept definition, the steps of of a procedure or rule, or the statement of a principle. The student is given FEEDBACK about the correctness of his answer.

EXAMPLES:

1. *"The father of our country is ________________?" (Fact)*
2. *"List in order the steps of procedure for ...." (Procedure)*

PRACTICE USING Component: The student is asked to use a concept definition, procedure, rule, or principle on a specific case to which it applies, and is given FEEDBACK about the quality of his performance.

EXAMPLES:

1. *"Classify the following Lofargrams." (concept)*
2. *"Using the procedure in the tech. manual, dissassemble the ...." (procedure)*
3. *"Solve the following circuit problems...." (rule)*
4. *"Predict the effect (sociological and psychological) when women are assigned to Navy ships." (principle).*

For CONSISTENCY, different components are required for different task levels:

For the REMEMBER level:	a STATEMENT	*(no example)*	PRACTICE REMEMBERING.
For the USE-UNAIDED level:	a STATEMENT *(or a review of the statement.)*	EXAMPLES *(at least one).*	PRACTICE USING.
For the USE-AIDED level:	*(The aid takes the place of the statement.)*	EXAMPLES WITH AID.	PRACTICE USING WITH AID.

These required components apply across all content types (facts, concepts, procedures, rules, and principles) for REMEMBERING, and all except facts for USING. For example, if the objective and test item called for the student to remember a fact, then the instruction must contain a statement of the fact to be remembered, and at least one practice-remembering item with feedback. No example is required, because it would be redundant with the statement.

CONSISTENCY also requires that each required component meet the following criteria:

1. STATEMENTS must be COMPLETE.
2. EXAMPLES must show application of the complete content.
3. EXAMPLES must match or reflect the conditions and standards required of the objective and the test as closely as possible.
4. PRACTICE must include FEEDBACK.
5. PRACTICE must be of the same task/content level as the test item and objective.
6. PRACTICE must match or reflect the conditions and standards required of the objective and the test as closely as possible, or be designed to help the student gradually learn the final task.

Most of the requirements above are probably obvious, but some are complicated. COMPLETENESS, for example, requires different prescriptions for different content types:

> *For a CONCEPT: "complete" means that all the critical characteristics of the concept, and their combination, are given.*
>
> *For a PROCEDURE: "complete" means that all the steps of the procedure are given in the proper order.*
>
> *For a RULE: "complete" means that all the steps of the rule are given in the proper order.*
>
> *For a PRINCIPLE: "complete" means that all the pre- and post-conditions, actions, processes, causes, effects, and results are stated, and the relationship between them is clearly stated.*

PRESENTATION ADEQUACY

Once all the required instructional components are present, and each of these components meets all of the consistency criteria, the ADEQUACY *of the presentation can be assessed. This is done by checking each instructional component (statement, examples, practice) for certain characteristics.*

A STATEMENT is ADEQUATE if it meets the following criteria:

1. The statement must be SEPARATED from the rest of the instruction. *This helps the student find the main idea. When the statement is separated, the key points stand out, and are not buried in the presentation. There are several ways to accomplish this goal:*

    a. *Set off the statement with boxes.*
    b. *Use a different color.*
    c. *Use a different type, or underline.*
    d. *Place on a separate page, or in a special place on the page.*
    e. *For audio or movies, pause before giving the statement.*

2. The statement must be IDENTIFIED. *After the statement is separated, the student should be told what it is. This permits the student's attention to be focused on the key points and their application, rather than the student trying to become generally familiar with everything in the instruction. One way to identify a statement is to use the word "statement." Other more content-oriented words are even more helpful:*

    definition *procedure for ____* *the principle of ____*

    Main Idea: *Key Point:* *General rule:*

EXAMPLE:

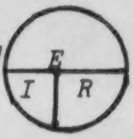
DEFINITION OF OHM'S LAW:
... ...

(Here, the statement is separated by the box, and identified.)

3. In addition to the statement, the presentation should include something to help the student better understand and remember the statement. *Methods of providing this help include:*

    a. *Giving a MNEMONIC (memory trick).*
    b. *Giving a general example of how the statement can be used.*
    c. *Explaining why the statement is important.*
    d. *Explaining how it came about, how it fits in the course, or how it relates to something the student already knows.*
    e. *Explaining some of the terms in the statement.*
    f. *Representing the statement with pictures, symbols, flowcharts, tables, etc.*

EXAMPLE: The following figure can be a helpful memory device for Ohm's. It will help you remember it so you can use it later on.



EXAMPLES are ADEQUATE if they meet the following criteria:

1. EXAMPLES must be SEPARATED and IDENTIFIED.

2. EXAMPLES must include some type of help.

3. EXAMPLES should range from "easy" to "hard."

4. EXAMPLES should be representative of the job the student will do after training.

5. There should be enough examples to cover the content area adequately.

6. EXAMPLES should clearly show why common errors are wrong.

The criteria are generally selp-explanatory. SEPARATED and IDENTIFIED are the same as for statements, and points 3 to 6 need no further explanation. The second criterion, HELP, is applied in different ways for different content types. Some types of HELP for each content type are given below:

HELPS for CONCEPTS: *Highlight the critical characteristics of an example.*
Explain why or why not something is classified as a member of a concept.
Show the use of a checklist or heuristic to help classify.
Simplify early examples, e.g. use line-drawings instead of complicated photographs.

HELPS for PROCEDURES or RULES: *Explain why each step is done.*
Explain why each step is important.
Give additional information about how to perform the task.
Give additional information about how to know if you've done it wrong.
Give flowcharts, tables, etc.

HELPS for PRINCIPLES: *Highlight important features.*
Simplify the relevant information from the case study in which it is embedded.
Use logical representations of the IF-THEN relationships.
Give additional information about how the principle applies, or why it doesn't.
Give hints as to how to analyze problems.

PRACTICE items are ADEQUATE if they meet the following criteria:

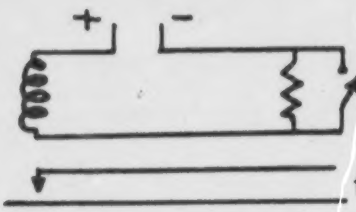
1. The PRACTICE section must be SEPARATE and IDENTIFIED.
2. The PRACTICE items must be free of hints that wouldn't be present in the test or on the job.
3. The PRACTICE items should have the same format as the format of the test items.
4. The PRACTICE items should range from easy to hard.
5. The PRACTICE items should be typical of the job to be performed after training.
6. The PRACTICE items should include the opportunity for common errors.
7. The FEEDBACK must also be SEPARATED and IDENTIFIED for each practice item.
8. The FEEDBACK should include help (similar to that for examples). *(As a bare minimum, the FEEDBACK should direct the student back to where the instruction was originally presented. However, it is better to have a new brief presentation, because if the student got the practice wrong, the original presentation didn't help enough.)*

The criteria are also self-explanatory.

EXAMPLE: *The next example shows an instructional presentation which violates many of the adequacy criteria described above. This example is followed by a more adequate presentation of the same subject matter.*

INADEQUATE PRESENTATION on the principles of operation of an alarm circuit:

The alarm circuit senses extremely high temperatures. When an extreme steam temperature occurs (which is a very dangerous condition that may have adverse consequences for a ship and her crew), the sensing switch contacts close thus shunting the resistor. The decreased resistance in the circuit, according to OHM'S LAW (E=IR), causes an increase in current flow in the circuit, which is enough to operate the alarm relay. The relay is designed to operate at a current flow above that normally found in the circuit. OHM'S LAW states that with voltage constant, a decrease in resistance in the circuit must be accompanied by an increase in current flow. The contacts of the alarm relay then close to actuate the audible alarm device, which may consist of a warning bell with an electrically operated clapper, or an H254 resonated horn assembly. Both of these produce extremely loud signals so they can overcome normal ambient noise levels.



to external audible alarm device.

Why is it important that the alarm circuit be operational at all times? Remember what hot steam can do to ships and sailors.

The example above is inadequate in several ways. First, the principle of operation of the circuit is not separated or identified. How is the student to know what to learn from this presentation? Second, the presentation is cluttered with a lot of other "nice to know" information that really doesn't help. If helps were included, they should aid remembering or understanding the principles of operation of the circuit. Also, the practice is not separate or identified, there is no feedback, and the practice really has nothing to do with remembering the principle.

The next page shows a more adequate presentation.

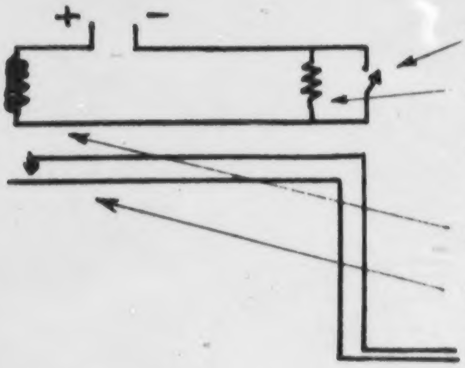
MORE ADEQUATE PRESENTATION on the priciples of operation of an alarm circuit:

OPERATION OF THE ALARM CIRCUIT:

(This section describes how the alarm circuit operates)

Extremely high steam temperatures cause the switch to close. This shunts the resistor, because the switch and the resistor are connected in parallel. Circuit resistance is decreased, and therefore, current flow is increased. The increased current flow operates the relay, closes its contacts, and energizes the bell or horn.

BASIC SCHEMATIC



EXPLANATION

1. High temperature closes switch.
2. Switch shunts resistor.
3. Decreased resistance = increased current flow. *(OHM'S LAW)*
4. Increased current operates relay.
5. Relay contacts close.
6. Relay contacts energize bell or horn.

PRACTICE: *Without using references or notes, explain how an alarm circuit operates. Be sure to include in your explanation the important actions that take place in the circuit. (Answer on pg. 256.)*

ANSWER TO PRACTICE QUESTION: *There are several ways you could have explained the operation of the alarm circuit, but your answer should have included the following ideas:*

1. *High temperature causes the switch to close.*
2. *When the switch closes it reduces total resistance in the circuit.*
3. *Decreased resistance means increased current flow.*
4. *The increased current flow operates the relay.*
5. *The relay contacts close and operate the bell or horn.*

USING THE IQI:

The IQI is designed for QUALITY CONTROL during any objectives-based instructional development process.

There are four documents that comprise the IQI:

1. Introduction and Overview *(This document)*
2. User's Manual *(contains all IQI procedures, and examples of their use)*
3. Workbook *(contains practice on all IQI procedures, with feedback)*
4. Job Performance Aid *(short version of all procedures)*

To facilitate use of the IQI procedures, the User's Manual, Workbook, and JPA were designed to include two quality control forms: The first form assesses objective adequacy, and the second is used to determine test consistency, test adequacy, objective-presentation consistency, and presentation adequacy. The suggested use of these forms is as follows:

1. Either during, or immediately after, the development of objectives in instructional development, use the objective adequacy form to assess the adequacy of each objective. Any required revisions should be made before instructional development proceeds.
2. As test items are developed for each objective, they should be checked for consistency with objectives, and adequacy, using the second form.
3. As new instructional materials are developed, or as existing materials are adopted, they are checked for consistency with objectives, and adequacy, using the second form. Required revisions to materials and tests are made before they are subjected to individual or small-group try-outs.

FOR MORE INFORMATION:

For more information on the IQI, contact: phone:

Navy Personnel Research and Development Center (714) 225-7121
Code P304
San Diego, CA 92152 AUTOVON 933-7121
7140
7194

REFERENCES

Ellis, J. A., Wulfeck, W. H. II, Merrill, M. D., Richards, R. E., Schmidt, R. V., & Wood, N. D. Interim training manual for the Instructional Quality Inventory (NPRDC Tech. Note 78-5). San Diego, CA.: 1978.

Merrill, M. D., Richards, R. E., Schmidt, R. V., & Wood, N. D. Interim training manual for the Instructional Strategy Diagnostic Profile (NPRDC Special Report 77-14). San Diego, CA.: 1977(a)

Merrill, M. D., Wood, N. D., Baker, M., Ellis, J. A., & Wulfeck, W. H. II Empirical validation of selected Instructional Strategy Diagnostic Profile prescriptions (NPRDC Tech. Report 77-43). San Diego, CA.: 1977(b).

Merrill, M. D., & Wood, N. D. Validation of the Instructional Strategy Diagnostic Profile: Empirical studies (NPRDC Tech. Report 77-25). San Diego, CA.: 1977.

Wood, N. D., Ellis, J. A., & Wulfeck, W. H. II Instructional Strategy Diagnostic Profile training manual: Workshop evaluation (NPRDC Special Report 78-17). San Diego, CA.: 1978.

DESIGN OF MACHINE SCORABLE "HANDS-ON" PERFORMANCE TESTS IN A PAPER AND PENCIL MODE

by

ROBERT N. JOHNSON
US Army Administration Center
Fort Benjamin Harrison, IN 46216

With the advent of job and task analysis in occupational training programs, the emphasis in measurement of student proficiency has shifted from the conventional multiple choice, paper and pencil test, to the performance test. Performance tests have obvious advantages. They require students to actually demonstrate their ability to perform job tasks to specified standards under conditions approximating a real world operational situation. Unfortunately, performance tests take more time and resources to administer, and are normally subject to inadvertent variation in the way they are administered and scored. These cost and reliability disadvantages are, in fact, the major advantages of the conventional multiple choice paper and pencil test. If the two design approaches could be combined, the resultant test might be called a machine scorable "Hands-On" Performance Test in a paper and pencil mode.

At the US Army Administration Center we have been experimenting with this type test for several years. To date results indicate that this type test is best used when three conditions are met.

1. The essential behaviors involved in the task to be tested are mental (or cognitive).

2. Task performance results in a tangible product with measurable characteristics.

3. The procedures or sequence used during task performance need not be measured during the test as long as the finished product meets specification (product measurement only).

With that introduction, let us move on to the test design rationale.

If we are to develop a machine scorable performance test, the designer must address five major considerations as shown in Figure 1.

CONSIDERATIONS

1. Conditions existing prior to task performance.
2. Initiating cues.
3. Actual task performance.
4. Results of task performance.
5. Cost effectiveness.

Figure 1.

The conditions existing aspect refers to the establishment and provision of an environment which will realistically simulate conditions which are similar to those under which the task is actually performed in the real world.

The second consideration refers to the necessity for considering the initiating cues that require the student to recognize the need to perform the task.

The third consideration includes the need to insure that each student will be required to actually perform all or most of the key elements of actual task performance during the test.

The results aspect refers to the need to insure that the test will serve two primary purposes. First, we must insure that the test results will separate those who can, from those who can't, adequately perform the task. Secondly, we must assure that test results will generate diagnostic information which can be used as basis for improvement of our training program.

And lastly we must provide an acceptable tradeoff between the efficiency of the test and the costs of test development and administration.

I will now address each of these considerations in terms of their inherent primary testing concerns in the design of a machine scorable performance test.

CONSIDERATION	PRIMARY TESTING CONCERN
1. Conditions existing prior to task performance	- Realism - Tools/reference availability Actual Simulated
2. Initiating cues	- High fidelity - Overcueing

Figure 2

With respect to conditions existing prior to task performance we are concerned with what we will provide the student to facilitate his performance during the test. The test situation should portray a realistic setting recognizable to the student as a feasible real world situation. It must also provide necessary background or peripheral information necessary to task performance but not included in the test requirement.

The test requirement should clearly state what the student is to do and what tools or references are available for use during the test. While realism is the key to the test situation, the test requirement may entail limitations to full fidelity performance. Tools and references may have to be simulated to facilitate a paper and pencil test.

Initiating cues must be examined in detail. If the task has straight forward initiating cues which are easily recognizable, the test requirement should do little more than state the performance requirement. In this case we would not be testing cue recognition. Many tasks, however, have single or multiple initiating cues which are difficult to isolate from other competency or extraneous cues which exist in an operational environment. When cue recognition is a major factor in failure to perform a task, care must be taken to insure that initiating cues are introduced in a high fidelity manner and that neither the test requirements nor the answer sheet overcue the student. In these cases the answer sheet must provide alternative responses based upon the real world behavior of failing to recognize the cue.

---

3. Actual Task Performance
   - Key Behaviors
   - Task Domain
   - Sampling From Domain
   - Task Integrity
   - Test Fidelity

Figure 3

Since both time and cost are major test design considerations, we cannot afford to test all behaviors inherent in an operational task. Accordingly, it is essential that the detailed task analysis be examined to identify the key behaviors involved in task performance. For linear tasks consisting of step by step procedures which are always performed in the prescribed sequence, the identification of key behaviors is no problem. (Example - Disassemble an M-16 Rifle). Most tasks, however, have a variable procedure during which steps in the procedure are not necessarily performed in sequence. The initiating cue or other cues generated during task performance dictate what is to be done next. In some iterations of these tasks certain steps are by-passed, in others the same steps may be repeated several times. An excellent example is the task of "computing a travel voucher." Most of us are here on government travel orders and will submit a travel voucher for payment upon return to our home stations. Some travel clerk will have to compute our vouchers. The initiating cue is generally the same; receipt of the voucher itself with supporting orders, receipts, itinerary, etc. How the task is performed, however, will vary as a result of the number and types of transportation used, types of expenses incurred, times of arrival and departure, etc. Each of these variables are, in fact, internal cues generated during any single repetition of the task which alters what must be done and how, for that specific case. These possible variations are dictated by the task domain. By task domain I am referring to the limits or scope of the task. We must analyze the task domain in terms of probable and possible variations of the task faced by the job incumbent in the field and then develop a rationale for sampling from that domain for testing purposes. Then, and only then, can we determine which of the key behaviors should and can be tested.

Once key behaviors are identified, the usual approach to paper and pencil machine scorable test design is to develop one or more items to address each separate behavior. We end up, therefore, with a test which measures component behaviors independently, without any assurance that the student can put them all together at the right time and place, and in the right sequence, to accomplish the task as a whole. This approach destroys test fidelity and integrity of the task, and results in a test of questionable face, content, or discriminate validity. If we are to assert that we have developed a true performance test we must assure that component behaviors are exercised by the student in the context of the total task much as he/she would in a real world environment.

---

4. Result of Task Performance
   - Mastery Standards
   - Training Feedback
   - Answer Sheet Design
     Realistic Alternatives
     Behavioral Alternatives
     Facilitates Error

Figure 4

With respect to results of task performance, we have several problems. Since we have limited the domain of required task performance, real world mastery standards may have to be adjusted. The trick is to establish test standards which separate performers from non-performers as defined by the student's ability to meet actual task standards on the job. Test validation procedures must address this primary concern. Since we are also concerned with training feedback, test designers must insure that each test not only properly identifies the non-performers but also facilitates identification of the cause of failure. When substantial percentages of students fail a test for identical reasons we have identified possible weaknesses or omissions in our training materials.

Answer sheet design is the key to training feedback. Over the years we have found that the use of real world alternatives facilitate training feedback analysis. In most cases the range of alternatives provided should encompass all real world behaviors which would be appropriate to any variation of the total task. The real world option of doing nothing must be included. By providing a complete behavior set in our alternatives, we allow the student to make errors of omission or commission. Any reasonable error will then be accommodated by an appropriate alternative. The alternatives available should never cue students to the fact that they have made an error.

---

5. Cost Effectiveness
   - Design Costs
   - Ease of Administration
   - Time to Administer
   - Validity
     - Content
     - Face
     - Discriminate

Figure 5

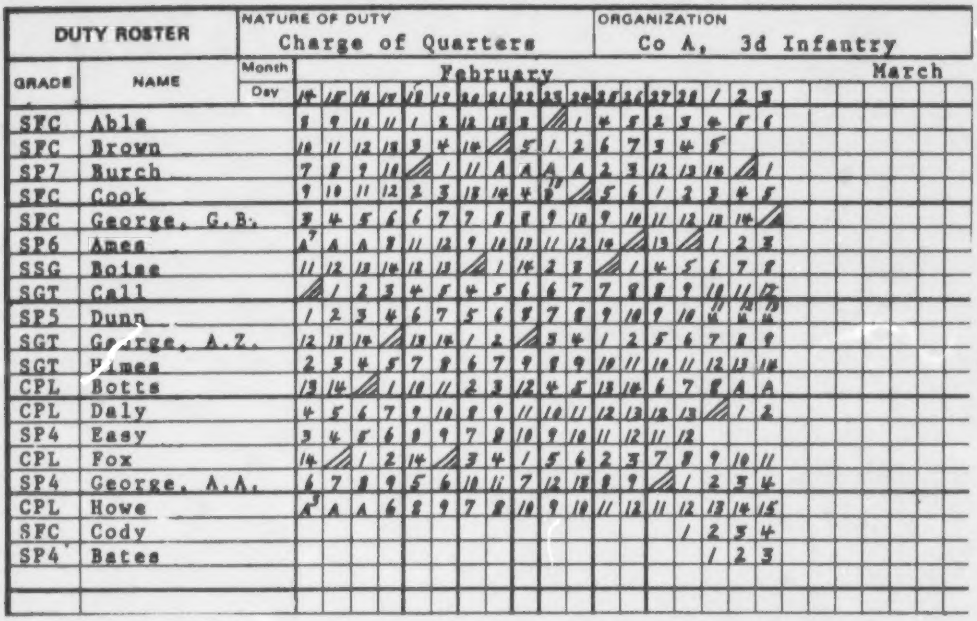
As with any test we must develop a balance between cost and effectiveness. Time or cost considerations may require reduction in the fidelity of the test, thus affecting its validity. If the level of test fidelity is lowered to the point where the student is no longer performing key behaviors in the context of the total task, we no longer have a "hands-on" performance test.

These then are the considerations, concerns, and principles we will use to develop a machine scorable "hands-on" performance test in a paper and pencil mode.

In order to demonstrate the application of these principles, I will use as an example the task "Select a Detail Using a Duty Roster." Since the full task analysis will be published in these conference proceedings (Figure 14) I will merely provide a task overview to facilitate understanding of the development process.

Every unit and most offices in the Army are tasked to provide a detail of one or more personnel to perform duties which are incidental to mission accomplishment. The duty roster provides a mechanism for spreading the burden equitably among eligible members of a unit. Normally, the senior NCO of each unit, or his clerk, maintains a duty roster for specific or general details. In the real world the unit is notified orally, in writing, or by Standing Operating Procedure to provide a specific number of personnel on a specific date for the detail. The task involves posting the current status of each eligible member of the unit on the duty roster in terms of availability or non-availability for the detail and determining who should be detailed based upon longest time since last selected. A short Army regulation dictates how availability is determined and selection is to be made but is not normally used during performance. The task requires personal knowledge or input as to the current and projected status of each member who may be suject to the detail. After posting the duty roster, the NCO selects and announces who will "pull the detail." The posted roster serves as the basis for the next iteration of the task. Improper selection or posting of the roster results in complaints from the troops and impacts on morale.

Just to keep you with me, I will now show you what a real duty roster looks like (Figure 6).



DUTY ROSTER

NATURE OF DUTY
Charge of Quarters

ORGANIZATION
Co A, 3d Infantry

GRADE	NAME	Month	February																					March		
		Day	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	1	2	3							
SFC	Able		8	9	10	11	1	2	12	13	3	▨	1	4	5	2	3	4	5	6							
SFC	Brown		10	11	12	13	3	4	14	▨	5	1	2	6	7	3	4	5									
SP7	Burch		7	8	9	10	▨	1	11	A	A	A	A	2	3	12	13	14	▨	1							
SFC	Cook		9	10	11	12	2	3	13	14	4	3^{11}	▨	5	6	1	2	3	4	5							
SFC	George, G.B.		3	4	5	6	6	7	7	8	8	9	10	9	10	11	12	13	14	▨							
SP6	Ames		A^7	A	A	8	11	12	9	10	13	11	12	14	▨	13	▨	1	2	3							
SSG	Boise		11	12	13	14	12	13	▨	1	14	2	3	▨	1	4	5	6	7	8							
SGT	Call		▨	1	2	3	4	5	4	5	6	6	7	7	8	8	9	10	11	12							
SP5	Dunn		1	2	3	4	6	7	5	6	8	7	8	9	10	9	10	11 14	12 14	13 14							
SGT	George, A.Z.		12	13	14	▨	13	14	1	2	▨	3	4	1	2	5	6	7	8	9							
SGT	Himes		2	3	4	5	7	8	6	7	9	8	9	10	11	10	11	12	13	14							
CPL	Botts		13	14	▨	1	10	11	2	3	12	4	5	13	14	6	7	8	A	A							
CPL	Daly		4	5	6	7	9	10	8	9	11	10	11	12	13	12	13	▨	1	2							
SP4	Easy		3	4	5	6	8	9	7	8	10	9	10	11	12	11	12										
CPL	Fox		14	▨	1	2	14	▨	3	4	1	5	6	2	3	7	8	9	10	11							
SP4	George, A.A.		6	7	8	9	5	6	10	11	7	12	13	8	9	▨	1	2	3	4							
CPL	Howe		A^3	A	A	6	8	9	7	8	10	9	10	11	12	11	12	13	14	15							
SFC	Cody																1	2	3	4							
SP4	Bates																	1	2	3							

Figure 6

Note that it contains nothing but personnel identification, numbers and letters which represent availability and selection priority and hash marks to indicate selection.

Let us now apply the principles outlined above to development of a machine scorable paper and pencil performance test.

With respect to providing realistic conditions we will provide a simulated duty roster correctly posted to include the last previous detail. Since current and projected status of each member is available on the job, we also provide this information. The reference regulation will not be provided. Policy on who is eligible for the detail, the date of the detail, the number of personnel required, and the current date will also be provided. Since the initiating cue is clear-cut, we will simply tell the student to post the roster and select the detail.

Accordingly, the instructions to the student would read something like this:

Test Situation: Your unit maintains a duty roster for police call. This is a weekly detail for which you provide one soldier for one day each week. All personnel grade E4 and below are eligible for this duty. Shown at Figure __ is the current status of the Police Call Roster showing the correct last column entry for the last detail on 10 June 1981. Figure __ also shows a note containing known personnel status changes as of today, 15 June 1981. You may assume that the status of each soldier remains the same as on the current roster unless the status notes indicate a change.

Test Requirement: Examine the Police Call Roster and read the notes contained in Figure __, then actually post the duty roster for 17 June 1981.

So far we created a realistic test situation and a "hands-on" performance requirement. Now we must develop the simulated duty roster and the status changes with which the student will work.

Our next step is to examine the task analysis and identify the key behaviors involved during the performance of the task. For this task the key behaviors are shown at Figure 7 (see next page).

KEY BEHAVIORS DURING PERFORMANCE

BASED ON CURRENT/PROJECTED STATUS:

A. Identify personnel to be added to roster.

B. Identify personnel to be deleted from roster.

C. Classify roster personnel as available or not available.

D. Classify non-availables into three categories.

  1. Authorized absence - (Code A)

  2. Other duty commitment (Code D)

  3. Unauthorized absence (Code U)

E. Advanced eligibility number by one for:

  1. Availables (number only)

  2. Code D non-availables

  3. Code U non-availables

F. Do not advance eligibility number for Code A non-availables.

G. Enter appropriate eligibility number, code, and/or name on duty roster.

H. Select most eligible available based upon highest eligibility number.

I. Select between equals by highest position on roster.

J. Erase entries for final selections and enter hash marks.

Figure 7

This is all easy enough. Now the problem is to identify the task domain and develop a reasonable sample thereof. There are many approaches but for this case we used a Test Content Matrix which contrasts what is to what could be or a before and after approach as shown at Figure 8.

TEST CONTENT MATRIX

AFTER →	SELECTED	A + #	D + #	U + #	# ONLY	NOT ON ROSTER
BEFORE ↓						
Selected						
A + #						
D + #						
U + #						
# Only						
Not on Roster						

Figure 8

Down the left axis we portray the situation which could exist on the current duty roster. Along the top we portray the situation which could exist after posting of the roster. The intersect boxes represent possible variations of the task or the total domain of task. For some tasks the elements on each axis may be different but, amazingly enough, they are often identical.

We now examine each intersect point and plot the key behaviors which would have to be applied to that specific combination. Once plotted, we review the results to see which variations require identical behaviors and whether all behaviors are included. Variations with identical behaviors are identified by a number representing that group of behaviors as shown in Figure 9.

BEHAVIORAL GROUPINGS

AFTER →	SELECTED	A + #	D + #	U + #	# ONLY	NOT ON ROSTER
BEFORE ↓						
Selected	N/A	2	3	4	5	6
A + #	1	2	3	4	5	6
D + #	1	2	3	4	5	6
U + #	1	2	3	4	5	6
# Only	1	2	3	4	5	6
Not on Roster	N/A	2 + 7	3 + 7	4 + 7	5 + 7	8

Figure 9

For example, each block or grouping labeled as number 1 requires the student to:

- Classify personnel as available.
- Advance eligibility by one
- Enter appropriate number.
- Select the individual with highest eligibility number.
- Erase number and enter hash marks.

To fully sample the complete domain of this task will therefore require at least one case for each of the behavioral groupings or a total of eight cases. To maintain realism, however, we need at least four cases of group 5 (availables) so that the student has a pool to select from. Accordingly, we need at least 11 cases to cover the waterfront. If test constraints preclude that number, some behavioral groupings can be dropped based upon importance. For example, group 8 may represent a rare and unusual circumstance which presents no real problem in the field. It would then be dropped. If the matrix fails to identify any key behavior, care must be taken to introduce it in its proper context. By this approach we develop a reasonable sample from the total domain of the task. What the student will see looks like Figure 10 (see next page).

DUTY ROSTER	NATURE OF DUTY Police Call	ORGANIZATION

GRADE	NAME	MONTH / DAY	June 10	17
SP4	Curtis		4	
SP4	Dickson		√//	
SP4	McManus		7	
SP4	Ramos		U1	
SP4	Steel		9	
SP4	Whitley		6	
PFC	Amos		5	
PFC	Barker		2	
PFC	Dunning		A3	
PFC	Johnson, C.		8	
PFC	Johnson, M.		8	
PFC	Turley		7	

Known Personnel Status as of
15 June 81

1. SP4 Curtis to confinement in hands of civil authorities as of 2200, 14 June.

2. SP4 Ramos returned from AWOL, 1800, 11 June.

3. SP4 Steel admitted to Post Hospital, 14 June. Line of Duty - Yes.

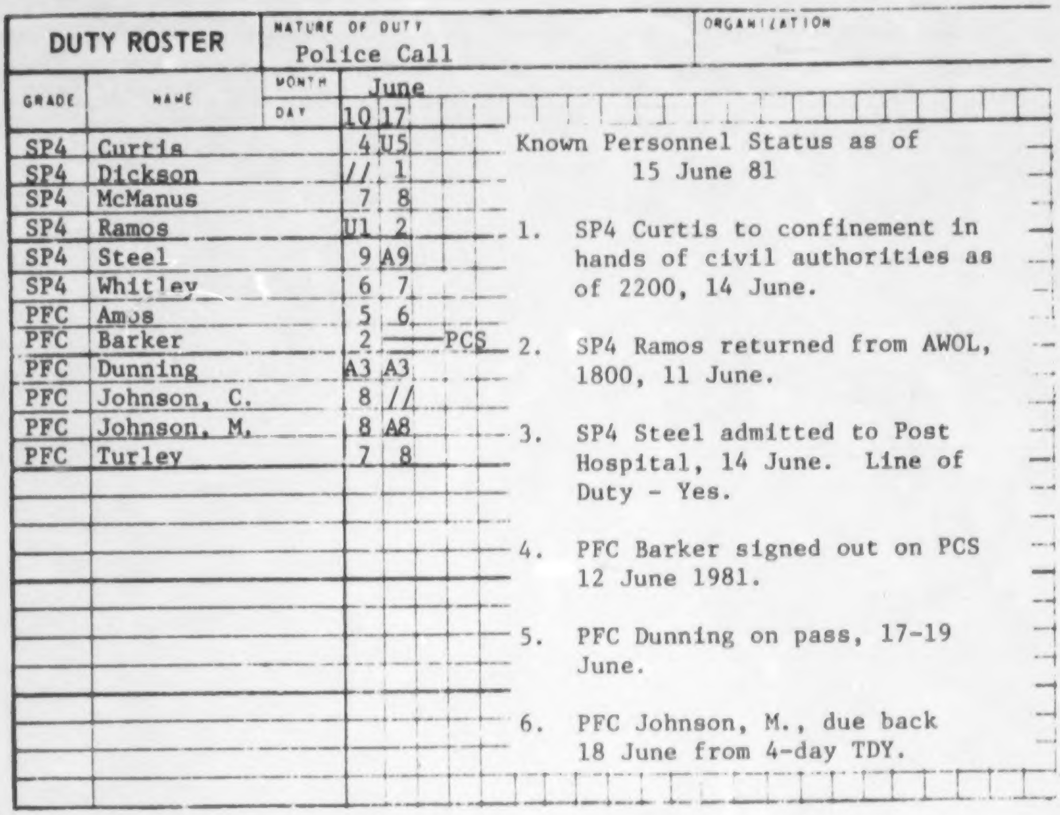
4. PFC Barker signed out on PCS, 12 June 1981.

5. PFC Dunning on pass, 17-19 June.

6. PFC Johnson, M., due back 18 June from 4-day TDY.

Figure 10

When he is finished posting the roster it should look like Figure 11 (see next page).



DUTY ROSTER

NATURE OF DUTY Police Call

ORGANIZATION

MONTH June

GRADE	NAME	DAY 10	17
SP4	Curtis	4	U5
SP4	Dickson	//	1
SP4	McManus	7	8
SP4	Ramos	U1	2
SP4	Steel	9	A9
SP4	Whitley	6	7
PFC	Amos	5	6
PFC	Barker	2	———PCS
PFC	Dunning	A3	A3
PFC	Johnson, C.	8	//
PFC	Johnson, M.	8	A8
PFC	Turley	7	8

Known Personnel Status as of
15 June 81

1. SP4 Curtis to confinement in hands of civil authorities as of 2200, 14 June.
2. SP4 Ramos returned from AWOL, 1800, 11 June.
3. SP4 Steel admitted to Post Hospital, 14 June. Line of Duty - Yes.
4. PFC Barker signed out on PCS 12 June 1981.
5. PFC Dunning on pass, 17-19 June.
6. PFC Johnson, M., due back 18 June from 4-day TDY.

Figure 11

So far we have created a high fidelity performance test in a paper and pencil mode but the result of performance must be hand scored. The problem now is to determine whether we can capture the essence of that performance in a student produced machine scorable format. We approach this problem by examining the final product itself and the task analysis to identify the characteristics of an acceptable product. For this task the key characteristics are shown at Figure 12.

PRODUCT CHARACTERISTICS

When appropriate displays a (an)

"A" (Authorized Absence)

"D" (Other Duties)

"U" (Unauthorized Absences)

Advanced Number

Unadvanced Number

Additional Names

Deletions from the Roster

Absence of an Entry

Hash Marks (Selections)

Figure 12

Regardless of the intermediate processes involved, the terminal behavior of the student is represented by these product characteristics. By restating these characteristics in terms of student behaviors we develop realistic and behavioral alternatives which can serve as the basis for our answer sheet as shown in Figure 13. Correct answers to the sample test are indicated on the answer sheet.

Using columns A thru d indicate how you posted your duty roster for the 10 individuals shown below for 17 June 1981 detail.

Reminder: More than one answer may be correct.

		A SELECTED FOR DUTY	B ENTERED "A"	C ENTERED "D"	D ENTERED "U"	E DID NOT ADVANCE THE NUMBER	F ADVANCED THE NUMBER	G ADDED TO ROSTER (ENTERED 1)	H DROPPED FROM ROSTER	I NO ENTRY MADE
000. SP4 Curtis	1	☐	☐	☐	☒	☐	☒	☐	☐	☐
000. SP4 Dickson	2	☐	☐	☐	☐	☐	☒	☐	☐	☐
000. SP4 McManus	3	☐	☐	☐	☐	☐	☒	☐	☐	☐
000. SP4 Ramos	4	☐	☐	☐	☐	☐	☒	☐	☐	☐
000. SP4 Steel	5	☐	☒	☐	☐	☒	☐	☐	☐	☐
000. PFC Amos	6	☐	☐	☐	☐	☐	☒	☐	☐	☐
000. PFC Barker	7	☐	☐	☐	☐	☐	☐	☐	☒	☐
000. PFC Dunning	8	☐	☒	☐	☐	☒	☐	☐	☐	☐
000. PFC Johnson, C.	9	☒	☐	☐	☐	☐	☐	☐	☐	☐
000. PFC Johnson, M.	10	☐	☒	☐	☐	☒	☐	☐	☐	☐
	11	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 13

In effect we now have the student actually posting the duty roster and then telling us what he has done in a machine scorable format. Note that the alternatives are a complete behavioral set. There is nothing else reasonable that the job incumbent can do.

The fact that two answer categories are not used as correct responses in this version of the test is irrelevant because those behaviors are feasible and will be used by non-performers who incorrectly perform the task. We are therefore facilitating error and avoiding overcueing by maintaining the entire behavioral set of alternatives.

A major advantage of an answer sheet with behavioral alternatives is that the test situation, test requirement, and answer sheet need never be changed. An infinite number of different tests can now be developed by merely changing the current and projected status portrayed.

We now must validate the test by administering it to a group of masters and non-masters to insure that it actually discriminates between the two. The validation will also help us to identify the cut score for this test which equates to full task mastery.

Note that the only unrealistic behavior required by this test is to transfer the actual coding to the answer sheet. This is considered worthwhile in terms of reduced costs of hand scoring and the generation of diagnostic training feedback. It does, however, produce an additional dimension to the validation procedure. The actual posting of the duty roster must be compared with the answer sheet during validation to identify the propensity for transcription error. If the training materials are designed with the same answer sheet, this problem normally disappears.

After administration, summarized test results in terms of item analysis will identify behavioral errors made by individuals or groups, thus facilitating the identification of weaknesses or omissions in our training materials.

By applying the principles and procedures outlined in this paper we can create "Hands-On" Performance Tests in a paper and pencil mode. Task integrity and test fidelity are maintained. Content, face and discriminate validity are inherent. Most importantly, the test does separate the men from the boys and provides detailed training feedback.

NOTE: The sample test displayed in this paper does not exactly match the rationale for selection from the task domain discussed on page 10. That test will be used in the near future and could not be compromised. The sample test displayed merely illustrates the approach.

TASK STRUCTURE ANALYSIS	1. DATE	2. PAGE 1 OF 4 PAGES
3. JOB/MOS: MOS: VARIABLE	4. DUTY: First Sergeant	
5. TASK STATEMENT: (STATE AS AN ACTION VERB WITH AN OBJECT.) Select A Detail Using A Duty Roster (DA Form 6)		6. TASK NUMBER

7. CUES: (LIST EVENTS WHICH INITIATE TASK PERFORMANCE.)	GO TO STEP	NECESSARY CONDITIONS
1. Oral or written requirement to select a recurring detail.	1	#, Date & Type of Detail.
2. Recurring requirement for a detail (SOP).	1	
3. Change in status of anyone on duty roster after publica-	30	Notification of Change in Status.

8. DECISIONS AND/OR STEPS: (STATE DECISIONS AS YES/NO QUESTIONS.) (STATE STEPS AS SUBTASKS.)	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
	YES	NO		
1. Do you have a duty roster for this detail?	2	25	X	All Active Duty Rosters.
2. Secure Duty Roster.	X	X	3	Appropriate Duty Roster, in Files in Office.
3. Are all columns of the roster already used?	25	4	X	
4. Annotate date of detail to next open column	X	X	5	
5. Are there any personnel to be added to the roster?	23	6	X	Notes Indicating Required Additions
a. New arrivals (ASGD or ATCHD).				Notification of Release from ED.
b. Permanent release from ED.				
6. Are there any personnel to be deleted from the roster?	24	7	X	Notes Indicating Required Deletions
a. Departures (reassignment or Rel from ATCHD).				Notification of New Ed.
b. New permanent ED.				
7. Any authorized non-availables? (LV, PASS, SD, TDY, SICK-LINE of Duty).	8	9	X	Notes Indicating Status of Individuals.
8. Post "A" opposite name under date of detail.	X	X	9	

ADMINCEN FORM 1997, 1 JUL 75
(ATSG-TE)

PREVIOUS EDITIONS MAY BE USED

Figure 14

TASK STRUCTURE ANALYSIS	PAGE 2 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
		YES	NO		
9. Any unauthorized non-availables? (AWOL, SICK-NLD, Confinement, Arrest, other reason due to own misconduct).		10	11	X	Notes Indicating Status of Individuals.
10. Post "U" opposite name under date of detail.		X	X	11	
11. Any eligibles who cannot be selected due to previous detail or other duty?		12	13	X	Other Detail Rosters, Knowledge of Other Duty Requirements.
12. Post "D" opposite name under date of detail		X	X	13	
13. Is this a consolidated roster?		14	15	X	Consolidated or Non-consolidated Roster.
14. Select previous column (if available) pertaining to category of detail (weekend/holiday of weekday).		X	X	15	
15. Identify (next) highest number in the selected previous column (if available) without an "A", "U" or "D" under date of detail.		X	X	16	
16. Is there more than one Soldier with the same highest number?		17	19	X	
17. Does the (remaining) detail requirement equal or exceed those identified?		19	18	X	
18. Select sufficient individuals to fill detail requirement by going down from top of roster.		X	X	19	
19. Place hatched lines, in pencil, opposite selected name(s) under date of detail.		X	X	20	Pencil.
20. Are more individuals required to fill detail requirement?		15	21	X	
21. With the exception of those posted with "A", add 1 to previous column running total and post under date of detail (use red pencil for weekend/holiday columns on consolidated rosters).		X	X	22	Red Pencil. Black Pencil.
22. File Duty Roster and publish Detail Roster (separate tasks).		X	X	EOT	
23. Annotate name to bottom of roster and line out previous detail columns, annotate reason on reverse side.		X	X	6	

ADMINCEN FORM 1997-1, APR 76
(ATSG-TE)

TASK STRUCTURE ANALYSIS	PAGE 3 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
		YES	NO		
24. Delete name from roster and annotate reason on reverse side.		X	X	7	
25. Secure blank Duty Roster.		X	X	27	
27. Identify all eligibles for entry on roster.		X	X	28	Unit Roster or Previous Duty Roster (Filled).
28. Enter names on roster alphabetically by pay grade, listing rank (SFC, SP6, SSG, CPL, etc)		X	X	4	
29. Post changes to duty roster.		X	X	30	
30. Is there any change in status of selected individuals in the detail roster which could preclude their pulling the detail?		31	EOT	X	Notification of Change in Status of Individuals in Published Detail Roster.
31. Erase hatched lines pertaining to those individuals and post new status.		X	X	15	Eraser.

PRODUCT CHARACTERISTICS:

! 1. List of selected individuals for detail

   a. Proper number.

   b. Correct names.

2. Properly posted duty roster.

   a. Correct date of detail in column heading.

   b. "A" posted by appropriate name.

   c. "D" and correct number posted by appropriate name.

   d. "U" and correct number posted by appropriate name.

   e. Hatched lines by appropriate names.

   f. Correct numbers posted by all other names.

   g. Correct names added to roster.

   h. Correct names deleted from roster.

ADMINCEN FORM 1997-1, APR 76
(ATSG-TE)

TASK STRUCTURE ANALYSIS	PAGE 4 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
		YES	NO		
i. Proper annotations made on reverse of roster. j. Correct heading on new rosters. k. Personnel listed alphabetically by rank on new rosters. *l*. Red entries for weekend/holiday details on consolidated rosters.					

SECTION 15

PERFORMANCE FEEDBACK

A Learning - Receptive State as Induced by an Auditory Signal or Frequency Pulse

Raymond O. Waldkoetter, Ed.D. and John R. Milligan, Ph.D.
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 3066, Fort Sill, Oklahoma 73503

Twentieth Military Testing
Association Conference
30 Oct - 3 Nov 1978
Oklahoma City, Oklahoma

A LEARNING - RECEPTIVE STATE AS INDUCED BY AN AUDITORY SIGNAL OR FREQUENCY RULSE

Raymond O. Waldkoetter, Ed.D. and John R. Milligan, Ph.D.

US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 3066, Fort Sill, Oklahoma 73503

INTRODUCTION

Many instructional procedures and techniques are and have been developed to make learning more effective. From the introduction of the printed text teachers have expounded on techniques for getting the student or subject to more readily learn and recall the procedural transmission of text content and material. Holding the student's attention and perhaps arousing a little motivational commitment seem to still have a high degree of relevance and educational concern. After the printed text came the development of audio-visual techniques and programmed text content. Yet relatively few students appear to become so entranced with cognitive or non-cognitive skill learning that they will persist in spite of the lure of television and other recreational distractions.

It would seem that added emphasis on the intrinsic, self direction of students to find a learning state that is anticipating and basically stress free should succeed where the extrinsic, apparatus oriented approach has not. This is not to advocate that the many advantages of apparatus in teaching and education or training be discarded with the instructional materials so conscientiously developed. Rather that the student's perceptual awareness and dynamics for focusing attention be re-examined to deliberately establish what sort of intrapersonal responses to promising stimuli indicates a more persistent receptivity for learning and success in subsequent evaluation.

---

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

By now the question must explicitly occur as to what methods and techniques are possible to bring about the self directed and intrinsic motivation of the learner as evaluated subject? In the context of my paper's direction, there is the accepted condition that attention is given to any stimulus that will achieve an independent identity whether pressure, light, sound or pain. Such a stimulus does not need to be at a conscious level of awareness, but can exist merely as an unheard sound or even suggestion. The necessary complement to the stimulus then is obtaining the student's or subject's awareness that the stimulus is present and can be responded to along with other intended behavior for an expected result.

Once this association is accepted, then a highly structured mode of instructional communication is required to relate to the stimulus which is unique and produces a facilitating response. Much as with the electrical pulses going through the telephone lines the voice message is accepted in the electrical current and reproduced for the listening party without any attending behavior toward the actual electrical pulse but only to the voice message. Accordingly, if a state is induced by an auditory signal using a pulsed sound frequency to ready the student and increase relaxation, an attentive rhythm will possibly occur to maintain a passively focused awareness. That is, once a characteristic alpha brain wave is induced by a particular set of auditory stimuli that condition should continue or be reinforced while positively suggested (voice) content material is presented in an initial or retrieved context.

METHOD

Now should the hypothetical state come about where a learner could respond to an auditory signal, it is conceivable that a more receptive behavioral mode would follow with less anxiety and a positive expectation for acquiring new information or recalling that already stored. There are contradictory experiences in the use of auditory signals and the method for inducing alpha brain waves. Research has shown some favorable experience for using alpha waves with a positively correlated relationship between percentage of alpha and memory (Green, 1973). This obviously sets the stage for exploring the use of the alpha wave state to figure out how learning may or may not attain specifically designated objectives with complementing positive reinforcement techniques for learning and retention.

At the edge of conscious attentiveness the alpha and theta brain waves may occasionally both appear. The more prevalent alpha frequency is functionally apparent even when the student's eyes are open, if properly conditioned. Usually with full physical reality contact, in an operational mode, the beta frequency is dominant. When the alpha wave is maintained from 8 to 13 Hz (cycles per second) and occasionally dropping below the 8 into theta, the student and trainee can experience a more complete sense of relaxation with attending auditory awareness.

In this state the individual is usually described as less ego involved and less inclined to have the usual susceptibility to inhibitions or so-called learning blocks. Within this state decision-making effectiveness will follow in a divergent pattern instead of converging in that many thoughts or images pass through the level of awareness with no conscious attempt to analyze barriers or disciplines, intuitive impressions are given free rein, the generation of many ideas is encouraged, and evaluative criteria are used primarily to create a synthesis of material for new ideas (Dirkes, 1978).

Since learning and information acquisition seem to require more capabilities than strictly programmed instruction permit, the learning and decision-making state must furnish ample opportunities to explore and reinforce those ideas or relationships that lead to other testable patterns without being an end in themselves. A frontier is recognized herein as the imperative need to take fuller advantage of mental potential by searching out how the learning-receptive state is attained efficiently and what is the most productive way to use such passively focused awareness. It is granted that total reliance on the relaxed condition would invite diminishing returns if too much importance is accorded the instilling process without ever getting into applied execution of ideas or decisions.

Although the introduction of the Losanov (1975) educational methodology is reported to have successful results in Bulgaria and now through an Iowa State University adaptation (Prichard & Schuster, 1978), the attentiveness of the student/subject may fluctuate depending on the instructional mode and environmental controls. This evolving Suggestive-Accelerated Learning and Teaching (SALT) approach consists of inducing a relaxed and receptive cognitive state in the student by conscious suggestions and then presentation of the learning material in combination with background music (sound) frequencies. The pragmatic results of the Losanov method and the Americanized version are open to critical challenge in some respects. There is nevertheless a consistent record of repeated uses of the techniques under the method showing both a more attentive student adjustment and increased acquisition and retention in a shorter span of instruction. While a mix of audio-visual and even tactile stimuli are employed, the fundamental reference point is instructor voice or audio direction and evaluation.

Research in this area generally shows a deficiency primarily in terms of integration of component learning or training parts. Methodology advocated in this paper is to bring about the introduction of a consistent auditory signal stimulus with combined cognitive-emotional suggestions carrying tactical information and the use of performance-oriented biofeedback. Because auditory guidance or signal frequencies are in part established as a known stimulus, it is further postulated that learning-receptive states of consciousness can yield positive learning effects as pulsed frequencies are experienced and instruction is phased into the monaural or binaural delivery and correlated biofeedback assists in clarifying performance objectives.

As exposure to a recently patented auditory guidance system (Monroe, 1977) has shown a potentially feasible approach in sound stimulus experimentation, effort has been invested in exploring the applied functions of such a system. Analysis of this Auditory Guidance System (AGS) paradigm should attempt to cover the "unified technology" of sound induction, content material design, and measure relationships to training effectiveness, modes of learning expression and perception, and states of conscious awareness. The major objective, then, is to try to investigate "what" improved learning and operational behavior could demonstrate more effective individual control and linkage of thought, informational, and memory processes.

RESULTS

Previous results from research in the area of anxiety and learning have consistently shown important relationships between various levels of anxiety and effectiveness of training (Isen, Clark, Shalker, & Karp, 1978). Most instructional technology largely ignores this set of relationships and must obtain further special elaboration to devise real applications to surmount unidentified frustration obstacles (UFOs) in trying to increase learning rate and mastery of complex behavior. Removal of cognitive-emotional barriers to effective learning is closely related to anxiety levels and has been substantially surveyed to identify targets for perceptive changes in gaining learning efficiency (McGrath & Cohen, 1978). Much of these research results have centered around building a learner's self confidence and receptivity by use of conscious and unconscious suggestion administered under specific levels of learner anxiety levels. Also, relatively sophisticated biofeedback instrumentation must provide verified relationships reinforcing the learner's capability to consciously control certain cognitive and emotional states favorable to learning receptiveness (Barber, 1972).

One attempt at a "unified technology" to change learning perceptions and responses, as illustrated by the SALT programs, strives to adapt knowledge from any pertinent field to accelerate the learning process by integrating cognitive-emotional stimuli into instructional programs. Conscious suggestions are given in the context of rhythmic performance with the background sound and altered modes of auditory expression and directed skill participation, reinforcing continually the feelings and attitude of relaxation and full satisfaction in performing the activity. Of many examples, both remedial work in language (Prichard & Taylor, 1976; Caskey, 1976) and teaching a junior high school science class (Gritton &

Benitez-Bordon, 1976) have led to significant positive results in surmounting past barriers and acquiring new information.

A slightly similar development which had its origin in the transcendental meditation (TM) movement and then broke away is that known under "the relaxation response" technique. Peters and Benson (1978) have reported highly positive results taking "the relaxation response" into a business setting from their Harvard research development site. They have provided consultative direction for voluntary "relaxation response breaks" resulting in significantly positive employee ratings of stress symptom reduction, improved performance and sociability-satisfaction.

Again, there is the recurring trend that physiological and psychological measures are strongly related and subject self control brings enhanced behavior and performance. Perhaps the remaining challenge is to discover how to precisely integrate the sound based instructions and rhythmic pulsing with properly reinforced learning modules and spaced training phases for performance skills.

In 1960 (Berlyne) a report of a Russian investigation described how pairing a tone signal with an electrical shock brought about a blood pressure or stress change. Gradually, though, with continuing trials of signal and shock in close sequence the response was extinguished, just as one usually adapts to a stimulus causing a mild irritant. However, with a change in stimulus pattern the tone by itself again evoked the stress arousal much as though one might respond to a cry in the night but only briefly attend when performing the multitude of concurrent day-time activities. Certainly learning and retention are effected by auditory stimuli to a recognizable degree. So, if the type of signal is available to induce and sustain a steady state of relaxed awareness with even possible peaked levels, and incisive, suggestively adapted course material is presented in well-focused, varied patterns, there should be a reasonable probability that both general and specific performance results are well within the scope of audio-guided behavior.

The AGS research being described by Monroe (1978) and demonstrated in stress-reduction workshops has identified a principal component in creating a newly innovated technique referred to as the frequency following response (FFR). There are cumulative experimental data showing how subjects respond to such sound frequencies structured to enhance the alpha brain waves and other psychophysical states. Such sound which moves through audible ranges also has masked pulses triggering what is termed the FFR. That is, there is synchronization of the signal and subject brain waves bringing a relaxed state, audible sound of surf and wind in the background, and the preparatory stage is set for altering alpha with programmed training modules and biofeedback monitoring. Drawing upon prior audiogenic discoveries and mnemonic instructional states, attention and learning dimensions can be charted based on the audio signals and combined voice instructions carried by mixed rhythms of monaural and binaural stimuli.

Following this line of exploratory development already verified in part by Monroe's generic patent of 1975, it is not inconceivable that research will quickly extend to take advantage of these partially confirmed audiogenic and adaptive listening pattern correlates. Adaptive learning behavior will build on a progressive series of FFR tape recordings letting the student experience differing information acquisition and perceptual dimension states. Using an adaptive mix of complex audio patterns, rather than static audio frequencies, carefully synchronized verbal guidance will instruct that selective listening techniques be passively focused on critical information processing requirements.

This approach could include a fully "unified training technology" of complementary suggestive learning and teaching precepts adhering to an engineered human resource model of training with sound, tailored course modules, and evaluative procedures. A parallel monitoring of electrophysiological activity would record further audiometric responses to indicate learning changes in attentiveness and perceptual modes. The extent to which audio stimulation and guided instructional content enhance operator capability would seem to deserve intensive research for probable high risk results to increase human potential in controlling complex mental activities.

DISCUSSION

Should the development of an AGS for accelerated learning techniques and instructional system design prevail in the face of those advocating only extrinsic motivation, it appears possible to markedly modify training patterns, perceptual modes and temporal states. By enhancing thought and information processing, memory and recall of data, human factor variables should function more reliably for intra- and inter-system operations. Learner and operator functions can have defined training requirements with selected critical tasks identified for sequential stages of assessed proficiency. Concurrently, experimental steps would analyze the patently valid basis of the AGS to evaluate any constraints in terms of information input functions and human storage security. By designing given training objectives, students following a programmed AGS sequence would furnish those data indicating the extent of AGS improved behavioral dimensions and operational performance.

Again, taking advantage of the proprietary AGS monaural and binaural stimuli, work should explore the relative scope of decision-making requirements involving novel human factor responses and functions of adaptive conscious states and associated physiological mechanisms.

Naturally, there are many questions needing answers in this developmental research area, substantiating even more the need for this comprehensive research strategy which may bear some similarity to the initial space research requiring interdisciplinary coordination. Now a realistic, well-integrated approach toward conquering facets of human, inner-space factors can produce new educational and behavioral practices for efficient learning and self control.

Rapid development of interactive computer systems and biofeedback instrumentation mark another convergence of scientific advances making the state of the art ready for audio and video response modes. Students could operate interactively in the future so that computer assisted instruction and self control of physiological parameters are synthesized. Assuredly, audio conditioning and guidance research achievements are moving into applied stages on a series of fronts running from sleep induction, stress and pain reduction, through suggestive-relaxed training.

The "unified training technology" to optimize intrinsic learning procedures and extrinsic motivational packages with computer assisted dialogues must not look that far away, unless one insists on denying the information and technological explosion. Many agencies, individuals, and systems are confronted with the challenge to deliver the intrinsic educational technology that will herald optimal student responses, while from another extrinsic direction we are exhorted to utilize more of our brain capacity. For example, in the wake of this turmoil, this past year a policy analyst (Fletcher, 1978) for the Deputy Assistant Secretary for Education noted that education would be completely revolutionized once a method could evolve to enable a person to have memory recall on demand or at least the processes for insuring retrieval.

What does this all have to do with personnel system testing and evaluation? You may rightly wonder! The AGS can yield in this "imagined" scenario within five to ten years that instructional technology assuring student attentiveness and rapid mastery of given subject-matter content. Verifiable responses would have the computer video support of adaptive, tailored testing breakthroughs (Urry, 1977) significantly testing with fewer questions and for greater psychometric efficiency. The highly structured student input will relate to the tailored testing and information theory and to a greater extent close the loop on diagnosing and prescribing accelerated or remedial learning conditions. Individuals should have more personal control for recall of their self contained micro-universe of test responses and respond more appropriately to the macro-content and selection search for precisely tailored test questions.

Over the 1990 horizon we may surely find an audio-video display terminal and AGS embedded training modules, the student interactively linked with the computer, and a wide assortment of tailored tests.

Certain audio and video stimuli patterns will guide relaxed but intensive self-retrieval searches. Alternating test trials should surmount emotional or skill barriers with precisely designed test responses. Between trial interpretive and transitional phases will suggest further guided instruction to store responses correlated with key evaluation criteria pinpointed by tailored testing dialogues.

In closing, might it now be agreed that acquisition and retrieval of information is aided with stress reduction as indicated by numerous verified measuring procedures? An affirmative answer would obviously suggest that instructional, information processing, and evaluative technology should now have the necessary design to include those auditory stimuli which induce more affectively integrated and responsive behaviors.

REFERENCES

Barber, T. X. Suggested "hypnotic" behavior: The trance paradigm versus an alternative paradigm. In W. E. Fromm & R. E. Shor (Eds.), *Hypnosis: Research developments and perspectives*. Chicago: Aldine-Atherton, 1972.

Berlyne, D. E. *Conflict, arousal, and curiosity*. New York: McGraw-Hill, 1960.

Caskey, O. L. Suggestopedic research in Texas. *Journal of Suggestive-Accelerative Learning and Teaching*, 1976, 1, 350-359.

Dirkes, M. A. The role of divergent production in the learning process. *American Psychologist*, 1978, 33, 815-820.

Fletcher, J. L. *The outer limits of human educability: A prospectus for a research program*. Unpublished manuscript, Washington, D.C.: Department of Health, Education, and Welfare, 1977.

Green, E. Biofeedback for mind-body self-regulation. In D. Shapiro, T. X. Barber, L. V. DiCara, J. Kamiya, N. E. Miller, & J. Stoyva, (Eds.), *Biofeedback and self-control*. Chicago: Aldine Publishing Co., 1973.

Gritton, C. E., & Benitez - Bordon, R. Americanizing suggestopedia: A preliminary trial of a U.S. classroom. *Journal of Suggestive-Accelerative Learning and Teaching*, 1976, 1, 83-94.

Isen, A. M., Clark, M., Shalker, T. E. & Karp, L. Affect, accessibility of material in memory and behavior: A cognitive loop? *Journal of Personality and Social Psychology*, 1976, 36, 1-12.

Lozanov, G. *Suggestology and Suggestopedia*. New York: Gordon and Brock Science Publishers, 1975.

McGath, J. J., & Cohen, D. B. REM sleep facilitation of adaptive waking behavior: A review of the literature. *Psychological Bulletin*, 1978, 85, 24-57.

Monroe, R. A. *Monroe Auditory Guidance Systems*. Unpublished manuscript. Afton, VA, 1977.

Monroe, R. A. Personal Communication. St. Louis, MO, 1978.

Peters, R., & Benson H. Time out from tension. *Harvard Business Review*, 1978, 56, 120-124.

Prichard, A., & Schuster, D. Third International Conference of Suggestive-Accelerative Learning and Teaching and Suggestopedia. Ames, Iowa: Iowa State University, 1978.

Prichard, A., & Taylor, J. Adapting the Losanov method for remedial instruction. Journal of Suggestive-Accelerative Learning and Teaching, 1976, 1, 107-115.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

SECTION 16

PERFORMANCE MEASUREMENT

Complexity of Flight Path Data as an Index of Skill in Piloting Performances from a Flight Simulator Based Job-Sample Test

Brian D. Shipley, Jr.
U. S. Army Research Institute Field Unit
Fort Rucker, Alabama 36362

To apply a flight simulator based, job-sample testing method to the problem of selecting trainees for Army helicopter pilot training, it was essential to develop a set of valid, reliable, and informative indicators of performance skill. The job-sample test provides a comprehensive time history record of each performance on twelve simulator control and instrument variables and two measures of side task performance. The problem of this methodological investigation was to develop a procedure for reducing the resulting mass of time history data to a few meaningful indices of performance skill.

Measures from existing research were deemed inadequate because such measures were: (a) unlikely to have any definite theoretical relationship to specific piloting behaviors of interest in pilot trainee selection research, (b) unlikely to provide sufficiently reliable measures of individual differences, and (c) unlikely to employ defensible commanded performance values as measurement standards and tolerances without a detailed verification study. Consequentally, the approach in this investigation was to derive an index of performance skill from a theoretical model, to establish a statistical basis for the data reduction process, and to develop a context free scoring procedure.

To test the operational feasibility of the data reduction methods, a set of computer programs was developed to simulate selected aspects of time series data as they might appear in piloting performances. The computer programs were used to construct theoretical samples of piloting behavior in a set of time series data. The resulting time series were analyzed with specially developed computer programs. The output of the analysis was evaluated in terms of the degree to which it recovered the known patterns of variation inserted with the simulation program. The purpose of this paper is to describe the data reduction and scoring procedures and their supporting rationale. Some outcomes from the analyses of the simulated time series data are presented to illustrate the data reduction process.

The Measurement Problem

Operationally, the objective measurement of piloting performances consists of three major steps. First, an interval or event sampling

1250

procedure is used to obtain a comprehensive time series record of the performance. Second, the time series record is condensed into a set of summary statistics. Finally, the summary statistics are translated into values which indicate the level of excellence achieved in the performance. The major difficulty in this operational sequence has been the need for general procedures to accomplish the second and third steps.

The measurement problem exists because, historically, all three steps were integrated by an instructor or standardization pilot into a single procedure which yielded a performance rating as the final product. Until the development of flight simulators and inflight data recorders, the typical researcher had to rely on a qualified pilot to obtain his data. Although modern electronic technology has freed the researcher from dependence on the pilot for his data collection, there is still a need for algorithms and supporting software to accomplish the second and third steps of the measurement process.

Data Summarization

In solving the second step, data summarization, the standard approach has been the simple averaging of differences between observed and an ideal or standard performance. These average values are inadequate for many applications because they obliterate essential information about the pattern of deviations from standard in the performance (Knoop & Welde, 1974). Another deficiency of the averaging method is the need for a valid standard performance. Knoop and Welde discovered that textbook descriptions of several maneuvers were not accurately reflected in the performances of highly experienced pilots. Consequentally, it would be desirable to have a data summarization procedure that would be context free, i.e., independent of an ideal or standard performance, and which would capture the major features in any arbitrary performance. As described in the next section, an appropriate method of determining the proper degree of fit can be used with the method of polynomial regression to achieve part of this data summarization objective. The objective can be entirely satisfied if the polynomial analysis is extended by the methods of Fourier analysis when appropriate.

Polynomial Regression

The objective of the present data summarization process is to capture all the worthwhile information in a time-series record without redundancy. Operationally, the correspondence between information and variance can be exploited to achieve a part of the desired result (see

Hays, 1967 for one definition). An orthogonal polynomial routine with a stepwise, forward fitting recursively defined algorithm can be used as the method of extracting information (Conte & DeBoor, 1972). With an orthogonal method, it is reasonable to assess the percent of variance accounted for with each newly fitted term as one criterion to determine the optimum degree of fit (Seber, 1977).

One criterion for determining optimum degree of fit is that any term account for at least 2 percent of the total variance. Cohen (1978) suggests that a term which accounts for at least 2% of the variance can be expected to have practical value in prediction. With this criterion, the analysis proceeds until a given term fails to account for the minimum amount of variance. In applying the forward solution with arbitrary data, there is one cautionary note. Seber observes that it is possible for the variance associated with the odd numbered terms to be very small if the shape of the series is nearly symmetric (Figure 1), and the variance test may incorrectly fail under such circumstances. In this case, an odd/even test of the terms number in the analysis can determine the need to extend the analysis at least one additional step.

Aside from the cautionary note, the polynomial regression may not always extract all the worthwhile information in an arbitrary set of data because of the minimum variance criterion. If the residual variance is sufficiently large, it is possible that it still contains worthwhile information. The method of mean square successive differences (Bloomfield, 1976) can be applied as an alternative criterion to test the hypothesis of information in the residual variance. The method of mean square successive differences takes into account the fact that time series data, as from an aircraft's flight path, are likely to exhibit a significant degree of correlation between adjacent values (Figure 2).

The mean square successive difference is easily obtained at each step of the polynomial analysis by computing the squared difference between each successive adjacent value, summing the result and averageing the sum over the number of degrees of freedom, (n - 1). This value is then used with the residual variance to compute a standard normal deviate, z-score. If the resulting z-score fails to yield a significant difference, the residual data are considered to represent a randomly sampled distribution and the analysis is terminated. The analysis proceeds as long as the z-score indicates that the residual data contains information, i.e., the sequence is nonrandom because it is serially correlated.

Suppose the z-score indicates that the residual data contains worthwhile information but the percent of variance in the polynomial regression has fallen below the minimum criterion of 2%. The method of Fourier analysis may then be employed to extract information about any periodic

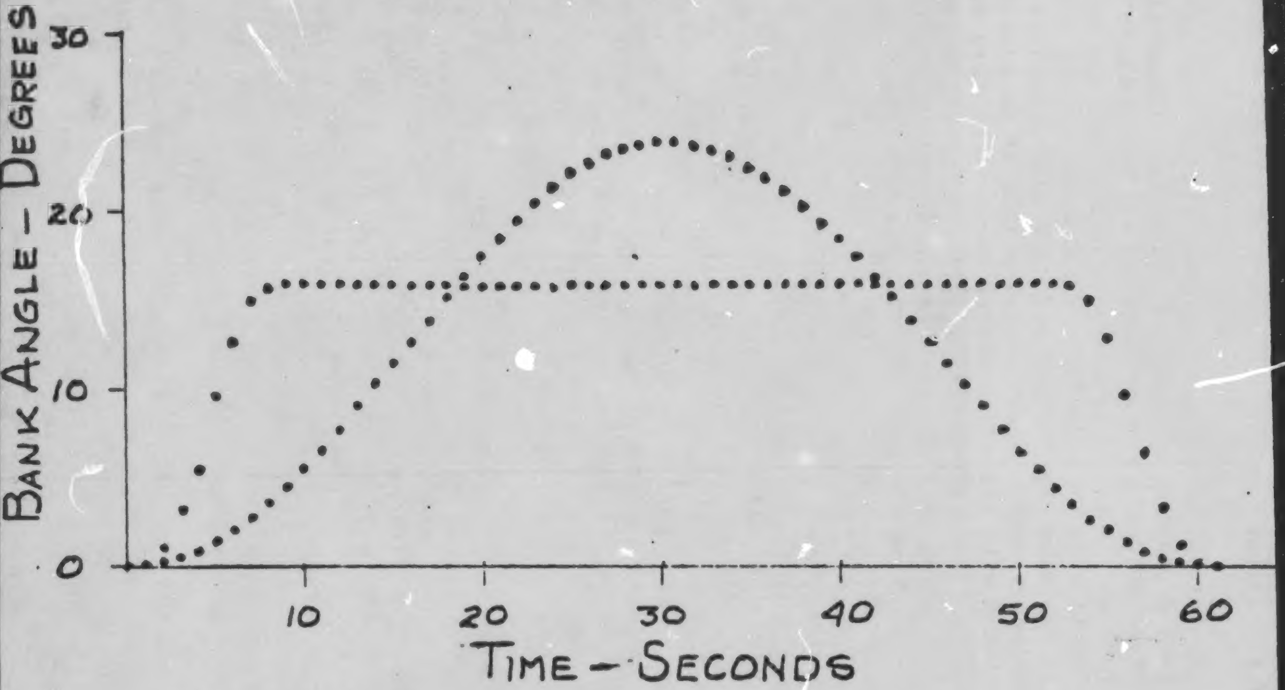
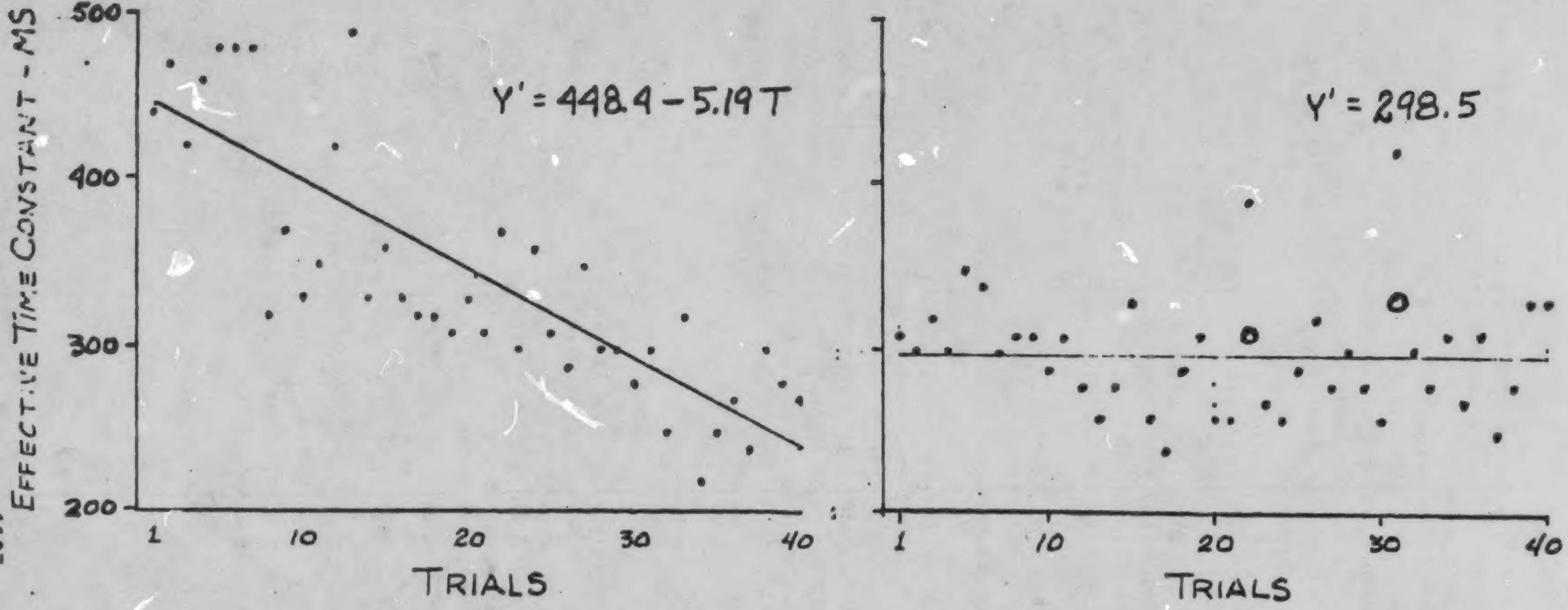


Figure 1: Two Examples of Symmetric Time Series Data



SERIAL CORRELATION

RANDOM SEQUENCE

FIGURE 2: TWO SAMPLES OF TIME-SERIES DATA EXHIBITING SERIAL CORRELATION AND RANDOM SEQUENCE - LEFT & RIGHT RESPECTIVELY

functions present in the residual data (Bloomfield, 1976). This method will extract short term periodic patterns in the data and as with polynomial regression, the Fourier coefficients are orthogonal. Because of orthogonality, analysis of variance methods can be employed to interpret the contribution of the periodic terms to the data analysis. In particular, each coefficient will account for some proportion of the residual variance. Thus, it is possible to apply the minimum variance criterion and a standard F-test to determine the practical value and the statistical significance of each Fourier term.

In summary, the output from this recommended data summarization process will be a sequence of polynomial coefficients with an associated proportion of variance. In addition, for some analyses there will be one or more significant coefficients for the corresponding periodic Fourier frequencies and each of these terms will also have its proportion of variance. This approach to data summarization has three distinctive virtues. One virtue is the capability to reconstruct the major features of the time series record from the given polynomial and Fourier coefficients, as is shown in Figure 3, i.e., the method captures the essential information in the data. With this reconstructive capability, it is unnecessary to retain the massive set of original data. A second virtue of the method is its ability to describe an arbitrary set of data, i.e., it is context free in that the data analyst is not required to postulate an ideal or standard performance in advance. For very long sequence of time series data, the analyst merely applies the methods of piecewise analysis by breaking the data into segments (Conte and DeBoor, 1972; Seber, 1977). Treated in greater detail in the next section, the third virtue of the method is the interpretability of the extracted variance patterns as an indicator of performance excellence.

Performance Evaluation

The third step of the measurement problem is to evaluate results from the data summarization as an indicator of performance excellence. While the interpretation procedure is easily described, its inherent value depends on an inference discussed in the next section about the source of variance in the time series data record. One interpretation method is simply to plot the variances associated with each coefficient in the polynomial and Fourier analyses in the order extracted by the analysis (Figure 4) or as a function of the Fourier frequencies (Figure 5). The problem is to determine the meaning of the patterns which might be exhibited by such a plot.

The concept of complexity is used here to evaluate the degree of excellence reflected in the pattern of plotted variances. Complexity, as used here, derives its meaning from a consideration of the internal

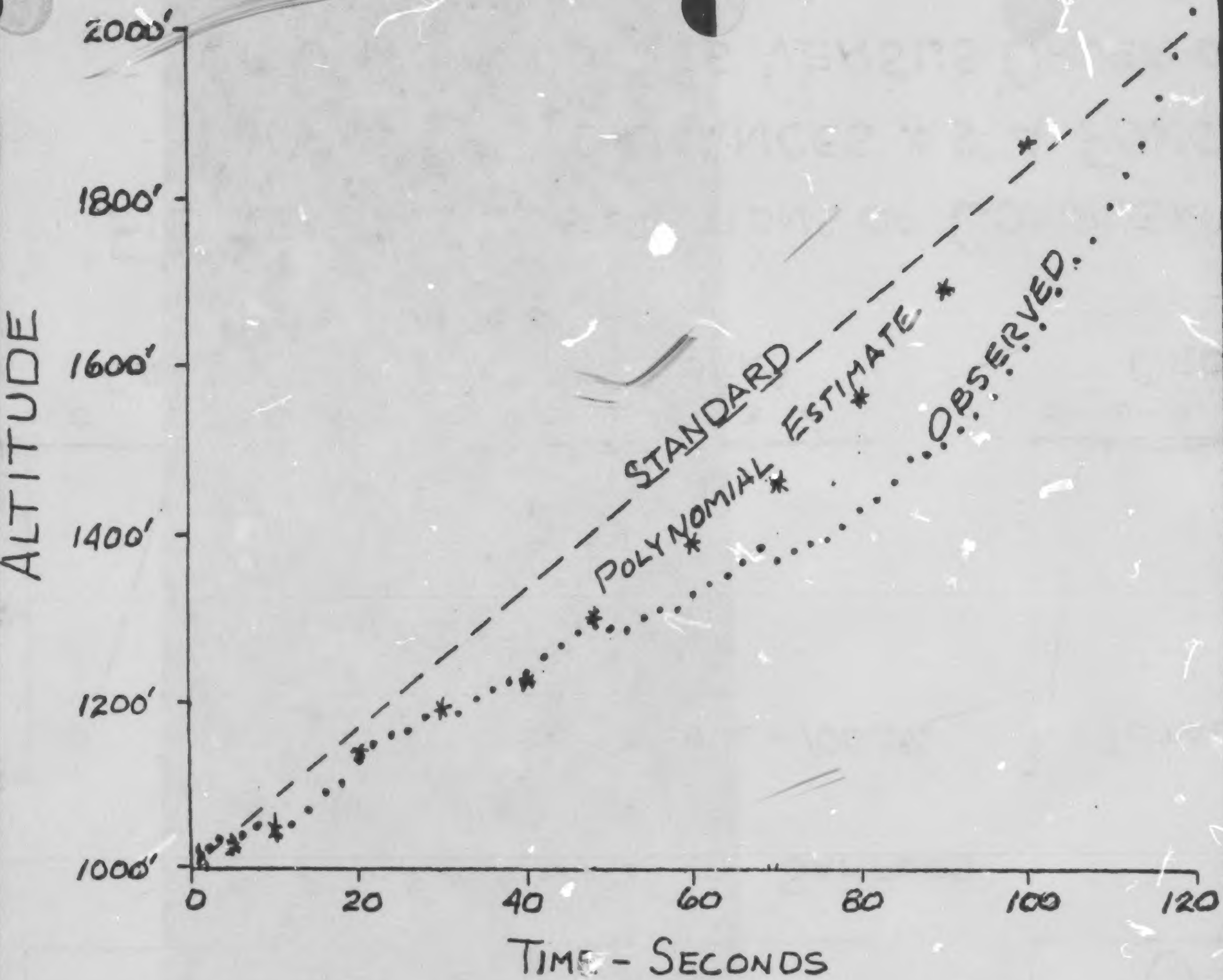
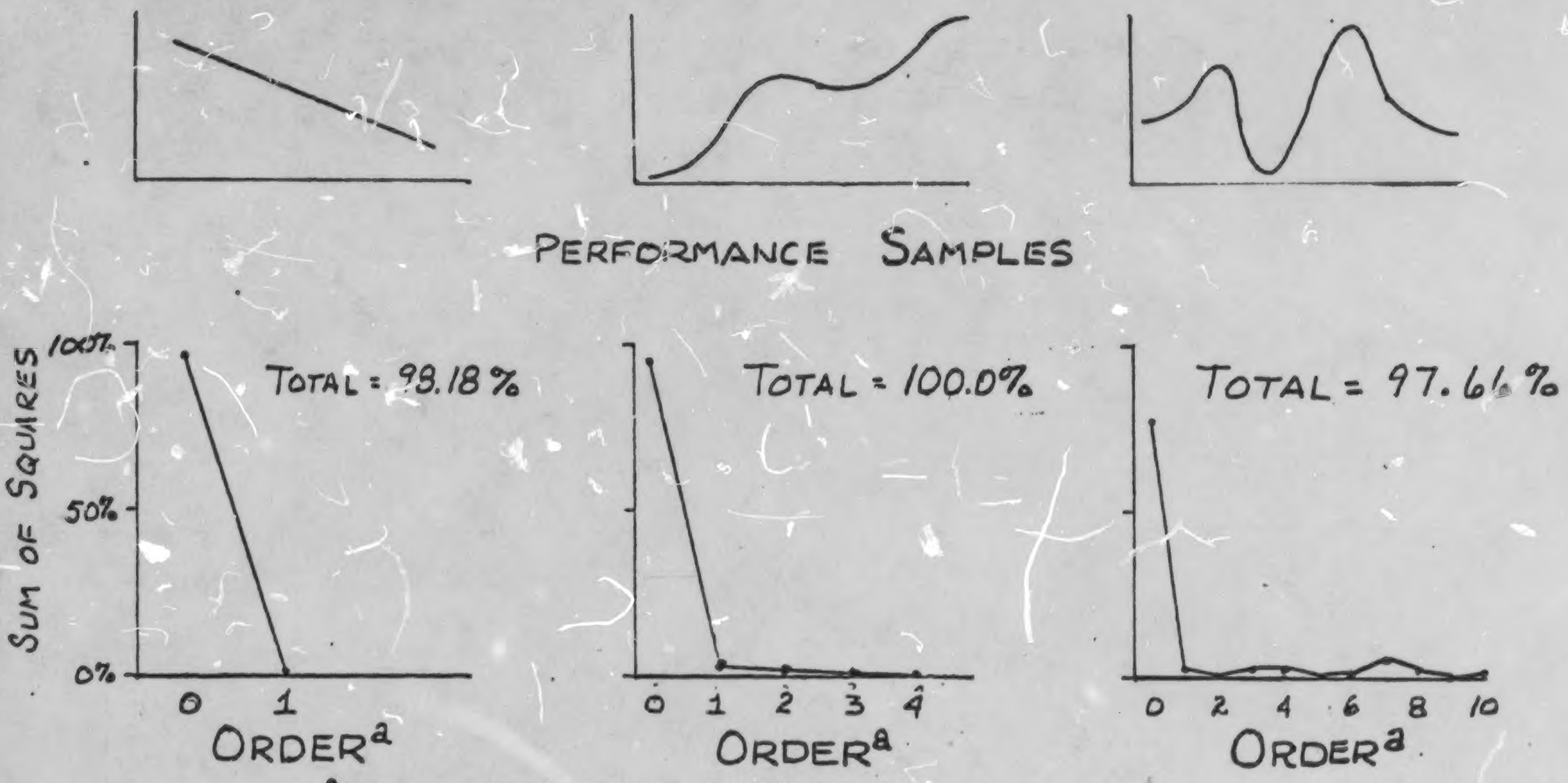


FIGURE 3: COMPARISON OF OBSERVED, POLYNOMIAL ESTIMATE AND STANDARD PERFORMANCES



^{a} NOTE CHANGE OF SCALE

FIGURE 4: ILLUSTRATIONS OF COMPLEXITY IN SAMPLE PERFORMANCES AS A FUNCTION OF SUM OF SQUARES VERSUS ORDER OF FIT

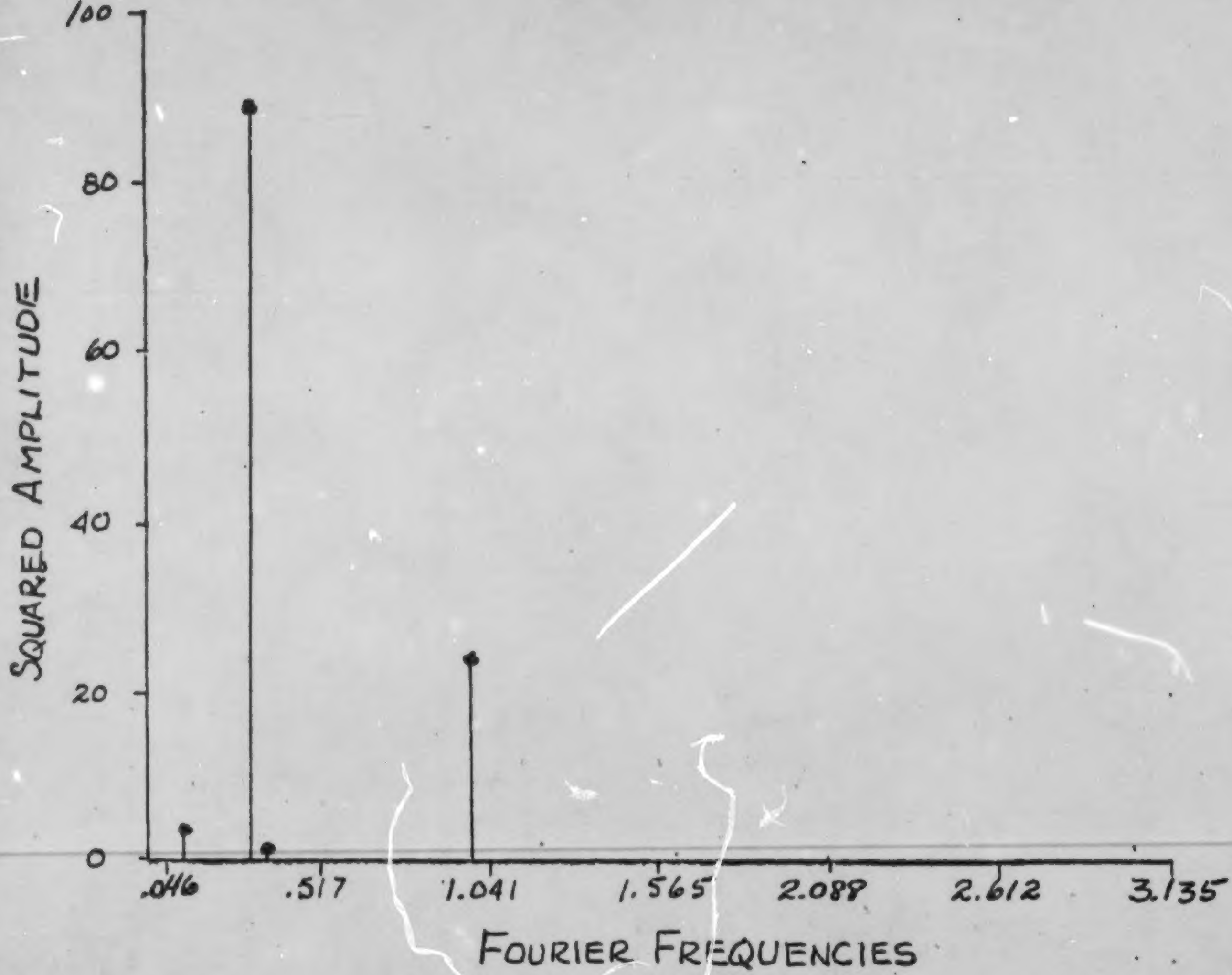


FIGURE 5 : PLOT OF NONZERO AMPLITUDES VERSUS FOURIER FREQUENCIES

structure of a set of time series data, e.g., an aircraft's flight path. Complexity is closely associated with the notion of serial correlation. A complex performance will exhibit relatively large deviations between adjacent or closely related values in time. In contrast, a simpler performance should have very few such deviations. By extension, it can be shown mathematically that a complex performance will exhibit more short-term trends and periodic terms than a simpler performance. (Bloomfield, 1976). It follows, then, that a complex performance should require more analytic terms to achieve a given level of information extraction.

The key to the translation of complexity from the plotted variances is the orthogonal basis of the data analysis. In the analysis, the base for determining the percent of variance is the total observed variability within that performance. In a forward analysis with orthogonal polynomials each successively fitted term accounts for proportionally less of the total variance (Seber, 1977). Consequentally, an analysis which requires many terms will, in general, reveal a smaller average variance over the number of terms fitted. Thus, number of terms and average percent variance should offer a useful index of degree of complexity. An objective interpretation of degree of complexity would employ a statistical modelling approach to fit the resultant plot of variances.

Complexity and Piloting Behavior

Level of experience, i.e., knowledge and skill in aircraft control can be linked to differences in degree of complexity. Kelley (1968) argues that the experienced pilot is readily able to convert his assigned mission into a projected flight profile, to anticipate the control movements needed to achieve that profile in a timely fashion, and to easily detect and correct minor errors of execution or random perturbation in aircraft performance. In short, it seems reasonable to characterize the aircraft control produced by an experienced pilot as generally smooth and regular, i.e., exhibiting a high degree of serial correlation, simple structure, and a low degree of complexity.

By contrast, the performance of the novice aviator should be rough and irregular, i.e., complex in structure. The novice has yet to acquire the necessary skills and knowledge associated with aircraft control. Unable to project the desired flight path sufficiently into the future, there will be many errors of omission. Unable to execute well integrated control movements, there will be many errors of commission. In short, the novice expends a great deal of time and energy attempting to dampen out his own errors, frequentally, without regard to the accomplishment of the overall objective. Nevertheless, as the novice gains experience, the resulting performances should exhibit a steady progression from greater complexity to greater simplicity as learning occurs. In a later report, data from a tryout experiment with the job-sample test will be used to evaluate the validity of the hypothesis that degree of complexity differentiates among performances of pilots at different levels of experience.

References

Bloomfield, P. Fourier analysis of time series: an introduction. New York: John Wiley & Sons, 1976.

Conte, S. D. & DeBoor, C. Elementary numerical analysis. New York: McGraw-Hill, 1972.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Kelley, C. R. Manual and automatic control: a theory of manual control and its application to manual and automatic systems. New York: John Wiley & Sons, 1968.

Knoop, P. A. & Welde, W. L. Automated pilot performance assessment in the T-37: a feasibility study (Tech. Rep. No. 76-6). Air Force Systems Command, Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, April 1973.

Seber, G. A. F. Linear regression analysis. New York: John Wiley & Sons, 1977.

EVALUATION OF INTELLIGENCE PRODUCING CAPABILITY OF SELECTED COMBAT ARMS UNITS

Earl W. Rubright

and

Alvaline Jackson

80th MTC

During 1977, without exception, units evaluated by the 80th Maneuver Training Command failed to demonstrate a capability to collect, synthesize or react to intelligence interjected into ARTEPs, CPXs or FTXs. This failing across all 39 Battalion size elements prompted the authors to purpose a system to measure the extent of this deficiency by looking at each MOS that influenced a unit's ability to develop combat intelligence.

The first task was to isolate where the problem lay. This meant structuring the intelligence information flow in a manner that permitted assignment of responsibility for collecting, processing and reporting a given piece of information. This chosen structure was based on the following data transfere points:

Indicator	Co. Level	Bn Cp	Bn S-2	Bn Co. or other Collection agency
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
I	II	III	IV	V

Data Transfere Points

Of necessity, the measures used were output not process. This meant the application of an absolute standard that was recognized by all concerned. For this standard two sources were used: Soldier Manual (SM) Tasks levels 1 through 5 for Enlisted Men and, for the Officers, the Infantry School Intelligence Training Objectives for Battalion S-2s. This single standard was used for Officers regardless of unit type or Officer Basic Branch.

Information inputs were selected from a pool of approximately 300 graphic and written Soviet force indicators. Each indicator supported a Soviet doctrinaire procedure and was aggregated to present current Soviet ground force doctrine. The decision to use this approach on how the intelligence picture should be acquired was based on the authors and USAINTS Collective Training Branch biases. Information from higher was expected to be sketchy, incomplete, and, if in written form, well nigh historical. The authors envisage an environment in which many of the technical collection capabilities will be severely impaired. The emphasis must be on combat intelligence and maximum utilization of organic resources, i.e., the troops. Combat intelligence is defined as an intelligence for the engagement of the moment. There are two aspects of information originating from below:

1. Where, in relation to your position, is it being obtained?
2. Is it a planned or accidental observation?

The prize goes to the S-2 that is acquiring his information well forward of his position through planned observation.

Prior to the exercises the OPORDs were reviewed to determine intelligence requirements. Once identified these requirements served as focus points that ensuing indicators clarified as the problem progressed. The clarity capable of being obtained by the S-2 depended on the relative importance of this intelligence requirement in relation to the unit mission. This was an arbitrary decision on the part of the evaluator.

How fast this clarity is obtained was dependent on the actions taken by the S-2, i.e., what he included in his collection plan. The first point is whether he developed enough information to determine when the Soviets would attack, where they would attack, and in what strength they would attack. The second and critical point is whether he told the Bn/Task Force Commander in time for the Commander to react.

The application of this to Battle Simulation at Bn and Bde level (Pegasus and CAMMS) FTX's, CPX'x, TEWT's and ARTEP's is as follows:

The S-2 of Bn/Bde in a defensive position receives an intelligence estimate. His responsibility, as defined, was to identify gaps in it according to his units mission and develop a collection plan that filled those gaps. Where he did this he received a comprehensive view of the enemy capability/disposition in the rear two-thirds (2/3) of Zone II. This, while not in strict accordance with the doctrinaire capabilities of his supporting intelligence collection agencies, forced Bn/Bde S-2s to depend on direct observation for the bulk of their information.

Once a determination was made as to what to portray based on OPFOR scenario, the problem was to select the series of indicators to use. Using US Army-Europe identification guides over 200 graphics for OPFOR vehicles and weaponry were developed. These graphics were deliberately made difficult to interpret by xeroxing and then presenting them at difficult angles for identification. As each problem progressed, each indicator became increasingly easier to identify through better reproduction and better angles of presentation. The graphics were used for actual observations, i.e., front line troops or RECON elements in an ARTEP, Company Commanders in a CPX.

In addition to the graphics each intelligence requirement was supported by a series of just over the horizon spot reports/indicators. Some of these spots were from refugees, tracer patterns, SAM launchings, noises, dust, detrious, flashes, shell holes, etc. Again, the emphasis was on the ability to identify and pass on significant information, whether it arrived at Battalion, and, finally, what happened to it at Battalion level. Additional indicators were abailable for technical reports from higher if requested as part of the S-2's collection plan, i.e., Side Looking Airborne Radar (SLAR) reports, USAF Infrared (IR) reports, ASA reports, etc.

This technique allowed the determination of the extent of any deficiency, where the deficiencies were located, and where the Commander needed to focus his training to overcome the deficiency. Through an informal arrangement with Individual Test Evaluation Directorate (ITED) of the US Army Training Support Center, where pertinent Soldier Manual (SM) skills are included as part of a

Skill Qualification Test (SQT), the Commander can also be told how his troops stack up in relation to his active Army counterpart.

As a review, the following sequence would be typical:

1845 hours: An OP spots two OPFOR Armored vehicles wheel out of and back into the opposite woodline. Total time of observation--3 seconds. What are the OP's responsibilities? The Infantry Soldier's Manual (FM 7-11) says he should be capable of:

1. SM Task 071-11A-0803 - Report activity using SALUTE
2. SM Task 071-11A-0806 - Identify Soviet vehicles
3. SM Task 071-11A-0802 - Speed captured documents to rear

This is the first Measurement Point, i.e., what is the OPs reaction and what does he report. In an ARTEP the Platoon or Company Evaluator/Controller actually shows a flash card to a troop in the field. In a CPX a board controller does the same to the Company Commander.

The second Measurement Point is what comes through to Battalion, i.e., to what extent does the Platoon leader and/or Company Commander act as an inhibitor to the flow of intelligence related information.

The spot report of the two armored vehicles upon arrival at Battalion is the responsibility of the Intelligence Sargeant who should be capable of:

1. SM Task 071-11B-8111 - Update enemy situation map
2. SM Task 071-11B-5430 - Maintain intelligence workbook

As these and other indicators come in they begin to form a picture of the intentions of the OPFOR. The next Measurement Point was what conclusions did the S-2 draw and what did he report to the Battalion or Task Force Commander. The original indicator was reintroduced at Battalion level if it was lost at any Measurement Point.

The use of this technique for information derived from subordinates measured several things:

In an ARTEP: 1. Did the troops demonstrate competency in SM Skills:

  a. SM Task 071-11A-0803 - Report activity using SALUTE
  b. SM Task 071-11A-0806 - Identify Soviet vehicles
  c. SM Task 071-11A-0802 - Speed captured documents to rear

In a CPX: 1. Did the Company Commanders demonstrate competency in SM skills:

  a. SM Task 071-11A-0803 - Report activity using SALUTE
  b. SM Task 071-11A-0806 - Identify Soviet vehicles
  c. SM Task 071-11A-0802 - Speed caputred documents to rear

ARTEP/CPX: 1. The extent to which the Company Commander inhibits the flow of information through his own ignorance.

2. Did the Bn/Bde Intelligence SArgeant demonstrate amony others compentency in SM skills:

  a. SM Task 071-11B-8111 - Update enemy situation map
  b. SM Task 071-11B-5430 - Maintain intelligence workbook

3. Did the S-2 correctly interpret the information provided?

4. Were the S-2 recommendations to the Commander timely and concrete?

For information derived from higher Headquarters, questions can be asked related to:

ARTEP/CPX: 1. Did the S-2 correctly identify gaps in the intelligence estimate when compared to his units mission?

2. Did S-2 provide guidance for development of collection plan?

3. Did the Intelligence Sargeant demonstrate competency in:

  a. SM Task 11B=5451 - Extract and use information from Intelligence estimate
  b. SM Task 11B-5470 - Prepare intelligence collection plan

c. SM Task 11B-5472 - Prepare Patrol Plan

d. SM Task 11B-5473 - Debrief patrols

In an ARTEP: 1. Did the Platoon Sargeant demonstrate competency in:

a. SM Task 11B-8305 - Plan and conduct zone Recon missions

b. SM Task 11B-8320 - Plan and conduct area Recon missions

It should be noted that these examples were only a portion of the required Solider's Manual Tasks reviewed.

The manning requirements for this technique were:

CPX: One individual with the Company Commanders to present graphic and written indicators and one person with the Bn/Bde S-2 to record what comes in and reintroduce indicators that fell through the cracks.

ARTEP: Platoon/Company evaluators were given packets of indicators to be presented to randomly selected troops between the hours of ____ to ____ on day ___ depending on the scenario. One person with the Bn/Bde S-2.

Because of the very limited number of scenarios addressed in the ARTEPs and battle simulations and the availability of an absolute standard for Enlisted Solider Manual tasks the validity of the technique was well grounded.

The design used was a compounded posttest only control group design. To illustrate this design graphically, the "R" represents random selection of units, "X" represents the administration of the ARTEP or Battle Simulation and "O" represents the administration of the indicator packets by the observer. It is important to remember that "X" and "O" are occuring simultaneously.

Step 1 consisted of stratified initial random selection of the first group, exposing this group to the X_1 (ARTEP/Battle Simulation) and the measuring of the results with the Criterion Referenced Test, O_1 (a packet of Soviet force indicators). This led to a composite design that could be graphically depicted as:

1270

Group 1 R X_1O_1

Group 2 R X_2O_2

Replication of Step 1 by unit 2 provided a control group for those unmodified portions of X_1 present in X_2.

A third unit was then selected and exposed to the indicator packets. This replication using nine (9) units led to a design that could graphically be portrayed as:

Group 1 X_1O_1

Group 2 X_2O_2

Group 3 X_3O_3

Group 9 X_9O_9

The single subject posttest design was selected by the authors for its appropriateness for the project and ease in administration. The model dealt with the following threats to internal validity:

1. History: History becomes a plausable rival hypothesis when specific events occuring between X_1O_1, X_2O_2, and X_3O_3 could be interpreted as causing a decrease in errors noted between O_1, O_2 . . .O_9. The probability that events that would influence population in all units in a similar manner and at the same point in time are remote.

2. Maturation: This can be ruled out because of the short period of time required to run the study for each unit.

3. Testing: This threat to internal validity can be ruled out as a possible rival hypothesis because retesting of the same unit did not occur. The reactive/obtrusive nature of the indicator packets was not controlled.

4. Statistical Regression: This would not become a rival hypothesis because the selection of the participants for each "X" and corresponding "O" was randomly made.

5. Instrumentation: This threat can be ruled out by the nature of the indicator packets.

6. Bias resulting from differential selection: This can be ruled out as a possible rival hypothesis because of randomly selecting the branch of the units.

7. Experimental mortality (loss of a portion of the experimental population) did not occur.

8. Since both selection and maturation can be ruled out, the threat posed by a selection maturation interaction can be disregarded.

The factors that are a threat to external validity are, unfortunately, not as easily dealt with.

1. The reactive or interaction effect of the packets was eliminated as a plausable rival hypothesis by the nature of the research design which did not call for a pretest. However, the obtrusive nature of the packet administrator may have had an unmeasured effect.

2. The interaction effects of selection and the experimental variable, i.e, the ARTEP/Battle Simulation cannot be ruled out. The use of an all male population as opposed to a male and female population could conceivably have biased the results. It was felt that the placement potential for females in the combat arms units below Battalion level did not justify the expense of including them.

3. Reactive effects of the experimental arrangement remains an open question. The limited population that was available necessitated the application of the indicator packets to all members of the population. Whether similar results would be obtained outside of the experimental setting is not known. The effect of the observer monitoring the operation was not measured.

4. The threat posed by multiple treatment interference is the most serious threat to this study. The use of a single population to assess the effectiveness of both Soviet defense and offense packets (the two packets that were used with

all units) was unfortunate but unavoidable due to the size of the available population. The exposure of the population to the same technique two or more times, i.e., the use of spot reports and/or technical reports in conjunction with pictures of each indicator series permits the possible interpretation of results on the Soviet defense packet being a function of exposure to the same technique of presentation that the unit experienced earlier with Soviet offensive packets. Other special packets (Airborne, AAA, River crossing, etc.) always followed these first two, if used at all.

Data Analysis

The results of the study were dealt with in two ways:

1. The average percentage of errors was computed for each unit
2. The average percentage of errors was computed for each SM Task or Intelligence Training Objective.

The measure of SM Task competency for levels 1-2 and for levels 4-5 presents an incredible picture.

For Skill Level 1-3

SM Task	Errors	No. of Possible Responses	Percentage of Errors
071-11A-0802 "Retrograde captured Documents"	5	11	45%
071-11A-0803 "SALUTE"	22	27	81%
071-11A-0806 "ID Soviet vehicles(RECON)"	23	27	85%
"1st Echelon Soviet"	23	27	85%
"1st Echelon WARSAW"	26	27	96%
"Equipment above Regiment"	25	27	92%
071-11A-0802 "OCOKA"	27	27	100%

$\overline{X} = 21.57$

$^{s}X = 7.52$

83.43 $\overline{X}$ error percentage

The nature of the deficiency, while well known, was not evisaged as being as acute as the data demonstrated

The problem, while statistically less acute for skill level 4-5, has greater potential for degrading the overall capability of the unit.

For Skill Level 4-5

SM Task	Errors	No. of Possible Responses	Percentage of Errors
071-11B-8111 "Update situation map"	1	7	14%
071-11B-8112 "Preparation of situation report"	0	7	0%
071-11B-8131 "Immediate air request"	6	7	85%
071-11B-5423 "Preparation of Overlays"	0	7	0%
071-11B-5430 "Establish and Maintain Intelligence workbooks"	6	7	85%
071-11B-5451 "Extract Intelligence Data from Intel Estimate"	7	7	100%
071-11B-5470 "Prepare Intelligence Plan"	6	7	85%

$\overline{X} = 3.71$

$s_X = 3.20$

52.71 $\overline{X}$ error percentage

The unfortunate conclusion is that only in mechanical skills can the NCO's at level 4-5 demonstrate any consistent proficiency. The skills related to individual interpretive capabilities and skills requiring the demonstration of initiative are those with the lowest scores. This seeming inability to act decisively in the absence of imaginative leadership came as a very real surprise. It was, frankly, a preconceived notion on the part of the authors that senior NCOs would take up the slack left by poorly trained company grade officers. This simply did not happen.

The presentation of the same requirement to troops in an ARTEP and Company Commanders in a Battle Simulation provided a unique opportunity to compare the capability of Company Commanders with that of their subordinates. Again, the results were disheartening.

	Errors	No. of Possible Responses	Percentage of Errors
Individual Soldier Response	34	36	94%
Company Commander Response	189	209	90%

The SM Task that requires the reporting of observed phenomenon (SM Task 071-11A-0803) showed the troops to be almost as proficient as the Company Commanders in including all aspects of the required communications.

	Errors	No. of Possible Responses	Percentage of Errors
Individual Soldier Response	4	6	66%
Company Commander Response	18	21	86%

If the only players in the game were individual soldiers, Company Commanders and Intelligence Sargeants then the picture would indeed be grim. Unfortunately, the compounding effect of ignorance does not stop here. The last link in the intelligence chain is the Bn S-2. To assess his ability to deal with adequate inputs from other members of the chain, the authors:

1. Reintroduced information that had been lost
2. Used the Intelligence Training Objectives of the Infantry School as criteria.

The results were appalling.

Intelligence Objectives	Errors	No. of Possible Responses	Percentage of Error
I 004 "Determine EEI"	6	7	86%
I 003 "Disseminate Combat Intelligence and Information"	7	7	100%
I 011 "Identify OPFOR actions through indicators"	6	7	86%
I 022 "Develop Collection Plan"	7	7	100%
I 029 "Analyze Doctrine and Tactics Employed by OPFOR"	6	7	86%
I 030 "Identify OPFOR support training weapons from Regiment down"	7	7	100%
I 033, 37, 38, 39, 48 "Identify sensors (any) required for mission"	5	7	71%
I 041 "Detect threats to Bn/TF Security"	7	7	100%

91.13 $\bar{X}$ error percentage

$\bar{X} = 6.31$

$^{s}X = .74$

As would be expected no generalizable finding across units were noted.

	Units 1	2	3	4	5	6	7	8	9	Respondent $\bar{X}$ % errors	Percentage errors across:
Soldiers				94	94					94	7 categories
Co. Cmdrs.	73	90	75			90	80	100	90	85	7 categories
Pltn. Sgts.				100	100					100	1 category
Intel. Sgts.	28	57	57			57	43	71	57	53	7 categories
Bn S-2	50	100	88			100	100	100	100	91	8 categories
Unit $\bar{X}$ % errors	53	82	73	97	97	82	74	90	82		

In all instances, the personnel being evaluated were given at least three opportunities to correctly identify, report, document or analyze the indicators. If they were successful on any of the three tries, they received full credit.

The inability to identify Soviet vehicles and weapons regardless of echelon was noted across all units for Company Commanders and troops. This data contrasts sharply with the first iteration of the Infantry SQT which indicated that 85% of all soldiers could distinguish Soviet combat vehicles. The ease with which Active Army troops seemed to pass the vehicle recognition requirement would lead us to question that particular portion of the SQT.

The percentages speak for themselves. Notably absent is the end result of so little expertise being demonstrated across so many differing skills. If the assumption is made that without these skills Bn/Bde size elements cannot produce combat intelligence then the probability of the success in a combat environment is very low indeed.

The essence of the process was repeated measures on the same information to determine if the information had been:

1. Recognized
2. Reported
3. Recorded/displayed
4. Interpreted
5. Used to generate request for additional information and/or recommendations for specific course of action on the part of the Battalion Commander.

For the U.S. Army Reserve the answer is, unequivocally, NO.

Learning Aptitude, Error Tolerance, and Achievement Level as Factors of Performance in a Visual-Tracking Task

Brian D. Shipley, Jr.
US Army Research Institute Field Unit
Fort Rucker, Alabama

INTRODUCTION

The Army Research Institute Field Unit at the Army Aviation Center is conducting aviator trainee selection research on job-sample, psychomotor, information processing, and time-sharing tests to improve the methods of selecting applicants for Army helicopter pilot training. This paper presents preliminary results from an investigation of methods to improve the measurement of visual tracking and time-sharing skill as a part of that research. In this section, the test is described, some sources of confounding are considered and methods to overcome the confounding are presented. Following the introduction, procedures are described for collecting data to test selected hypotheses about confounding. Then, the results of the data collection are presented and the discussion section focuses on the prospects for employing data from the visual tracking tests in time-sharing and aviator trainee selection research.

Visual Tracking Test

The visual tracking test used in the current research was designed to measure an individual's ability to control an unstable system. The test device is a single axis, compensatory visual tracking task described in Pew, Rollins, Adams and Gray (1977). The operator's task is to try to maintain a light spot in the center of a horizontal display using lateral movements of a finger operated joy-stick.

The test difficulty is controlled by the system time constant in the periodic processing of the control stick signal. The system time constant is a weighting function which determines the rate of change of light spot location in relation to control stick movements. The system time constant operates as a divisor so that the size of the constant is inversely related to test difficulty. The test device periodically samples the control stick signal and computes the location of the light spot as a weighted function of the present control input and a residual component from previous control signals added to the present light spot location value. The residual component is correlated with the operator's previous control behaviors and greatly increases the difficulty of learning effective control of the light spot.

The tracking test device can be operated in two difficulty modes: critical and fixed difficulty tracking. The fixed difficulty, or fixed tracking mode was designed primarily for time-sharing applications. In this mode, the tester fixes the time constant at a given value and the operator performs for a fixed period of time. The measure of skill in fixed tracking mode is the total absolute deviation of the light spot from the center of the display, averaged across the time of performance.

The critical difficulty, or critical tracking mode is used to estimate the operator's effective time delay. The effective time delay represents the minimum operator response time for the detection and correction of errors in continuous control tasks and is used as a parameter in human information processing and optimum control theory models of operator behavior. Operationally, the effective time delay is an index of the amount of time required for the operator to detect an error and to convert information about that error into a precise control movement. Estimates of the effective time delay from the critical tracking mode are employed as the value of the fixed time constant in the fixed tracking mode.

To measure the effective time delay in critical tracking mode, the test device progressively increases test difficulty as a function of time in the performance. Difficulty is progressively increased by systematically reducing the size of the time constant as a function of time in performance. As the time constant grows smaller, the rate of change in light spot location per unit time increases. Eventually, the rate of change in light spot location becomes so rapid that the operator is unable to maintain effective control, the location exceeds the limits of the display, and the performance ends. The measure of skill is the estimated effective time delay which is the size of the system time constant at the end of the performance. This investigation was designed to evaluate possible confounding effects in the measurement of critical tracking skill, i.e., measurement of the effective time delay.

Confounding Effects

A review of recent research with the present test (Pew et al., 1977) and two similar visual tracking tests (Damos, 1977; Gopher & North, 1974; North, 1977; North, Harris & Owens, 1978) suggested that the testing procedures had resulted in a confounding of other performance factors with the measurement of visual tracking skill. Pew et al. defended their procedures with evidence of test-retest reliability (Rose, 1974).

In the research with similar tests there was evidence that confounding effects had degraded the validity of the visual tracking data to estimate time-sharing capacity and would probably degrade the validity of these measures in aviator selection decisions. Gopher et al. (1974) and North (1977) observed improvements in time-sharing performance as contrasted with predictions from single-task performance. Gopher et al. offered three hypotheses which might account for these discrepancies: (a) Use of adaptive logic did not accurately estimate single-task tracking skill; (b) There was an improvement of single-task tracking skill as a function of practice in the time-sharing test; and (c) There is an independent time-sharing skill which is learned only in practice with time-sharing tests. At the conclusion of his report, North (1977) suggested that "isolation of improvement factors is an important direction for further research" (p. 92).

Two investigations addressed the question of confounding sources. In a transfer of training experiment, Damos (1977) found weak evidence of improvement of both single-task and time-sharing skill as a function of practice in multiple-task performance. Indications of confounding effects in the Damos (1977) data were: (a) operator unreliability as evidenced by heterogeniety of variance; and (b) failure of 16.7% of the subjects, 8 of 48, to achieve minimum criterion in subsequent time-sharing practice.

Although not specifically addressed by the authors, some difficulties with the use of adaptive logic to determine test difficulty were apparent in the investigation of test-retest reliability by North et al. (1978). The adaptive logic was used to establish tracking test difficulty in the first part of two daily testing sessions. After fixing the level of difficulty, the mean root-mean-square (RMS) tracking error was computed as the baseline for feedback on tracking performance in the time-sharing tests. Table 1 is a summary of correlations among the tracking task difficulty and RMS tracking error scores across the two daily sessions and two days of testing.

It is apparent from the data in Table 1 that test difficulty correlates negatively with dual-task RMS tracking error. This has potentially serious consequences in aviator trainee selection research because individuals who invest greater effort, and thus achieve higher levels of difficulty, would have greater difficulty demonstrating higher levels of time-sharing capacity. Conversely, individuals with low effort in the test difficulty phase would more easily exhibit greater capacity in time-sharing. In addition, Table 1 shows a significant decrease of correlation between single-task and time-sharing RMS error between the first and second days of testing. Since the high test-retest correlation (r = .90) between test difficulty across the two days of testing shows that the subjects were consistent in the amount of effort invested in the measurement of test difficulty, there were differential changes among individual RMS error performances as a function of changes in single-task performance. This is supported by the low reliability in single task RMS performance (rs = .01 & .34) and the moderate test-retest reliabilities of RMS dual-task performance (rs = .49 & .69).

Therefore, the available evidence suggests that procedures for measuring task difficulty allow for two major sources of confounding: (a) failure to train to asymptote before measuring single-task achievement, and (b) using current performance error as a criterion for adaptive adjustments of test difficulty. The first source of confounding could apparently be removed by training to asymptote or by developing a statistical model which accurately predicts asymptotic level of achievement from selected observations of learning performance. To remove the second source of confounding it was necessary to explain how differences in individual goals, effort, motivation and the like might interact with

Table 1

SELECTED INTERCORRELATIONS AND TEST-RETEST CORRELATIONS AMONG MEASURES OF TRACKING TASK DIFFICULTY, SINGLE- AND DUAL- TASK RMS TRACKING ERROR[a]

	Day 1 RMS Dual-Task	Day 2 RMS Dual-Task	Test/ Retest
Session A			
Task Difficulty	$-.53^b$	$-.43^b$	$.90^b$
RMS Single-Task	$.52^c$	$.13^c$	.01
RMS Dual-Task			.49
Session B			
RMS Single-Task	$.59^c$	$.10^c$	.34
RMS Dual-Task			.69

[a] North et al., (1978), p. 16

[b] Probability is less than .05 that the absolute value of any correlation greater than .388 is greater than zero; $\underline{t}(.388) = 2.064$, $\underline{df} = 24$.

[c] Probability is less than .05 that the differences between each pair of Day 1 minus Day 2 values is greater than zero; $\underline{Z}(.52) - \underline{Z}(.13) = 2.14$ (Fisher's $\underline{r}$ to $\underline{Z}$ transform).

single-task difficulty to obscure level of achievement and then to provide a means of measuring the degree of the interaction in an individual's tracking test data. As suggested in the following discussion, an adequate solution to the degree of effort problem is necessary to improve the validity of forced as well as adaptive difficulty testing paradigms.

Tolerance for Error

In a review of human performance limitations in visual tracking tasks, Poulton (1969) uses "tolerance for error" to explain how individual effort interacts with measures of tracking task ability. When first introduced to a relatively easy task, i.e., one with a single dimension or a simple control system, Poulton says that initially the operator will be challenged and interested in the task giving considerable attention and effort to task performance. Poulton continues:

> But...[the operator] soon discovers what he can and cannot achieve, and settles down to give what he considers to be an adequate performance. A small error comes to be tolerated, and effort is directed only at preventing or correcting large errors (Helson, 1949, p. 495). The task becomes analogous to a vigilance task, and fails to occupy the man's full channel capacity or attention.
>
> At this stage the level of performance can be improved by presenting the man with a challenge....knowledge of results can reduce the size of the error which the man will tolerate, and so raise the standard of his performance.
>
> Unfortunately, a change in experimental conditions that makes the task harder may also present a challenge to the man. This means that the poorer performance which is to be expected as a result of increased difficulty of the task may be partly offset by the challenge effect. Tracking in one dimension is thus not as sensitive to changes in experimental conditions as are tasks which occupy the man's channel capacity more fully... (1969, pp. 312-313)

Poulton's analysis indicates that the operator may decide to limit control effort to the prevention or correction of large errors. In his view, this decision converts the task from pure tracking to vigilance performance conditions. Success in vigilance performance is determined by error detection, the degree of error to be tolerated, and skill in error correction. Error detection will reflect differences in operator vigilance strategy. To prevent large errors, the operator maintains a higher level of attention or effort to anticipate and respond to performance conditions which, if uncorrected, would result in unacceptably large errors. On the other hand, when the operator strategy is to correct large errors, the operator responds only if he has detected the occurrence of deviations which have exceeded his acceptable tolerance limit.

An operator shift from pure tracking to one of the vigilance performance strategies would explain how the adaptive logic in the Gopher et al. (1974) testing paradigm allowed subjects to exhibit differential improvements over baseline predictions in dual-task performance. The adaptive logic in the Gopher et al. paradigm was expressed as a function of target error measured as deviation from center of the visual display. When error was consistently less than 10% of display length, task difficulty was progressively increased. If error consistently exceeded the 10% limit, task difficulty was reduced. Task difficulty stablized when the errors were distributed about equally above and below the limiting value. Given stable or increasing levels of skill, an operator decision to tolerate greater error would cause an increase in the observed deviations which would, in turn, cause a decrease in the existing estimate of task difficulty. The amount of decrease would be a direct function of the increase in error tolerance. In subsequent performances the operator would be able to achieve correspondingly less average error than predicted for higher levels of difficulty because the observed estimate of task difficulty underestimated the true level of skill.

Although the tolerance for error process invalidates existing procedures to estimate task difficulty with an adaptive logic approach, it must also be accounted for in a forced difficulty paradigm, e.g., Pew et al. (1977). Poulton's analysis implies that a decision to limit control effort represents the end of a learning phase in skill acquisition. However, the operator might become bored, fatigued, or otherwise disinclined to maintain effort to learn or perform before completely mastering the task. Estimates of task difficulty before a decision to switch from tracking to vigilance performance would thus underestimate the true asymptotic level of achievement. As an aside, there would be some training management value in knowing the extent of any skill improvement which might occur as a function of practice after the switch to the vigilance mode of performance.

The concept of tolerance for error and the corresponding switch from tracking to vigilance performance strategies has definite measurable implications. Suppose performance is represented as a sequence of observations of a measure of skill from repeated trials across some extended period of time. If greater effort in the learning phase corresponds to improvement of skill level and a constant or perhaps decreasing level of performance variability, data from the repeated observations should exhibit a definite trend of improvement of level of skill. An increased level of error after the shift to the vigilance phase should be observed as a discontinuity of either mean or variability of performance. In the vigilance phase, the observations should represent random samples from a distribution with mean and variance determined by the degree of error tolerance and the particular vigilance performance strategy. Statistical methods for estimating parameters from repeated observations will be considered after a brief summarization of the implication that an operator may attempt to minimize effort rather than maximize performance.

To summarize the implications of Poulton's concept of tolerance for error, it was hypothesized that (a) differences in operator goals, attitudes and the like would be represented in different performance strategies, (b) these strategies could be operationally defined on a scale of performance effort, and (c) different strategies and tolerances for error would lead to measurable differences in patterns of performance associated with the corresponding level of effort. The two extremes of the scale of effort would be performance maximization at the high effort end and effort minimization at the low end. Figure 1 depicts a schematic layout of the scale of effort concept and the ordering of performance strategies which were logically differentiated in the preceeding analysis of the tolerance for error concept.

The Method of Statistical Analysis

Standard statistical methods from the area of time-series data analysis provided the analytic tools needed to evaluate both trend and variability components in a sequence of tracking performance observations. Since these methods are commonly used in engineering and economic analyses, some of them may not be familiar to the psychologist. An understanding of mean square successive differences (MSSD) is crucial to the interpretation of the results of this investigation. Therefore, MSSD is described in limited detail here. Readers interested in greater detail should refer to the technical sources and those already familiar with MSSD may skip to the next section without any loss of continuity.

Mean square successive differences is a measure of variability of performance based on the order of the observations as the origin. As a measure of trend strength in a set of time-series data, e.g., repeated measures, MSSD derives its meaning from the fact that pairs of adjacent observations will be more highly correlated than will be pairs of more widely separated values. This sequential dependency of the observations on their order means that with a trend present in the data, differences between pairs of adjacent observations will be smaller than when the data is from a random sample. The variance is the average variability of the observations with the mean as the origin. Therefore, a comparison of the variance with MSSD will be an index of trend strength. When there is a linear or polynomial trend in the data, the MSSD will be small relative to the variance as illustrated in Figure 2. Without a stable trend, MSSD will approach the variance as a measure of variability. (See Brownlee, 1965, pp. 221-223 for a proof and more detail on computational methods.)

Standard methods are used to transform the ratio of MSSD to the variance into a standard normal deviate, i.e., a z-score (Brownlee, 1965). As a standard normal deviate this transformed ratio can be employed to

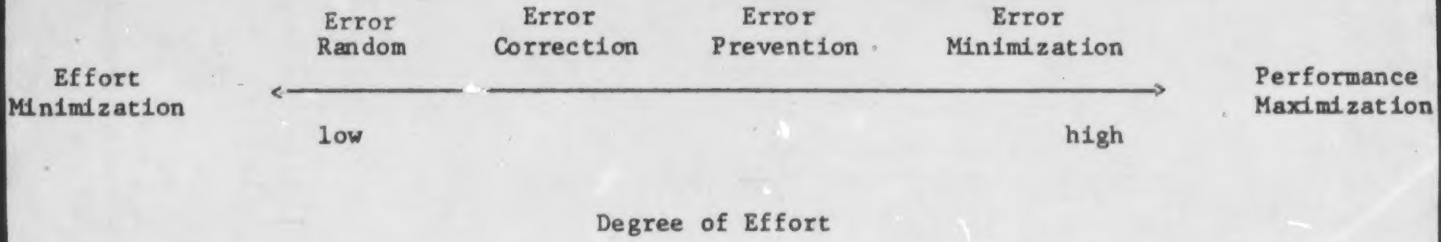

Error
Random
Error
Correction
Error
Prevention
Error
Minimization
Effort
Minimization
Performance
Maximization
low
high
Degree of Effort

Figure 1. Schematic Depiction of Tolerance for Error as a Function of Degree of Effort

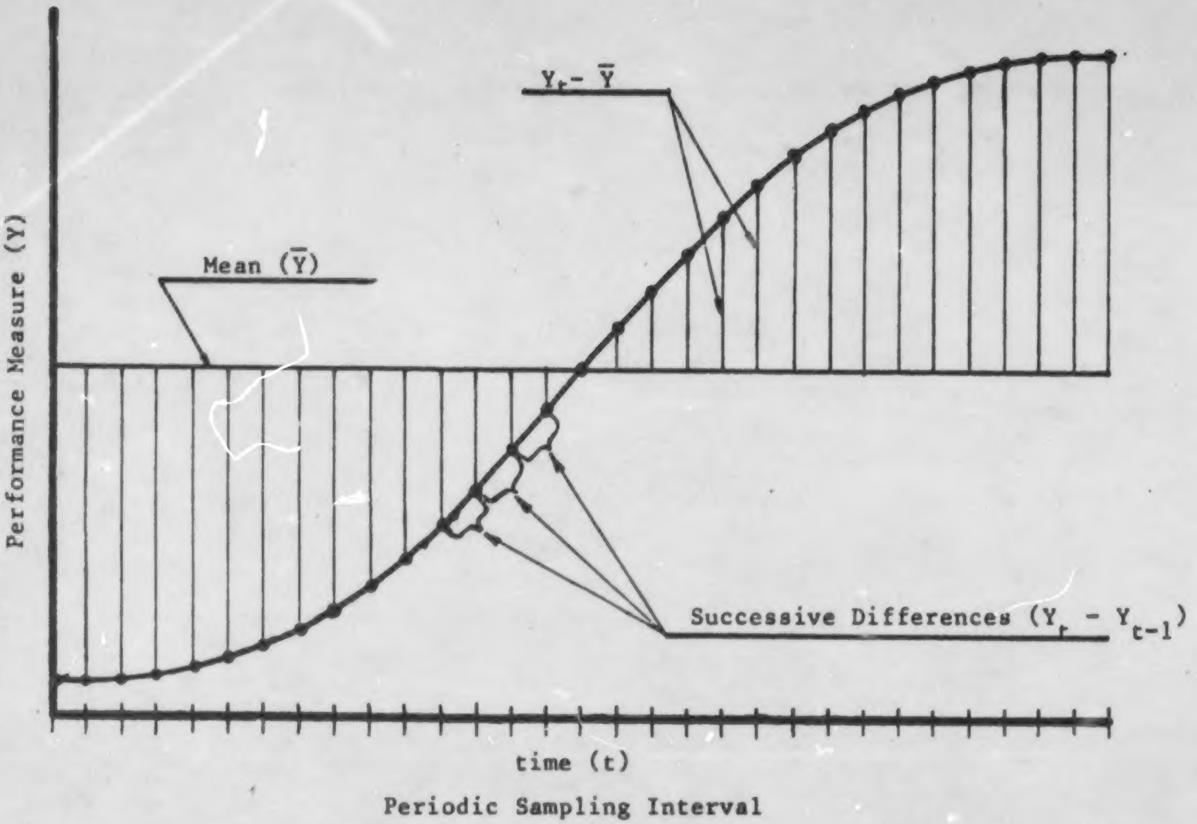

$Y_t - \bar{Y}$
Mean ($\bar{Y}$)
Performance Measure (Y)
Successive Differences ($Y_t - Y_{t-1}$)
time (t)
Periodic Sampling Interval

Figure 2: Depiction of relationship between successive differences and deviations from the mean in a set of time series data with a polynomial trend.

determine the departure of the data from randomness in the conventional statistical way. That is, the investigator posits an alpha probability and accepts or rejects the null hypothesis of no trend as the obtained z-score indicates. Brownlee reports that other investigators have shown that the z-score transform is acceptable with as few as ten observations and tables exist for use with as few as four observations. Unfortunately, these tables are not generally available and the occasional user may find it difficult to obtain copies (see Hart, 1942, for tables).

Research Hypothesis

The preceding analyses suggested that (a) the concept of tolerance for error would associate changes in performance effort and differences in such attitudinal variables as operator goals, motivation or interest in the task with differences in patterns of performance, particularly variability of performance, over time; and, (b) the MSSD measure would discriminate the presence or absence of trends in time series data. Suppose that two groups of subjects were selected on the basis of presumed differences in attitude, that if present, these attitudinal differences would result in differences in performance effort, and that members of these groups were given a series of trials with the Pew et al. (1977) visual tracking test in critical tracking mode. Finally, if the MSSD measure was then used to categorize performance by the members of each group into subgroups of random or non-random, analysis of trends or variability in the data for the resulting two by two contingency table should reveal an interaction of attitudinal group with type of performance across blocks of performance trials. The trials would be blocked to provide means and standard deviations to estimate the "local" level of achievement and variability of performance. The following data collection and analysis methods were employed to test this hypothesis of a triple interaction.

METHODS

Subjects

Data for this investigation were obtained from the records of 29 individuals who had participated in a comprehensive selection testing research program. Nine of the individuals had recently resigned or been eliminated from warrant officer or helicopter pilot training and 20 of their contemporaries were still in the Army warrant officer helicopter pilot training program at th US Army Aviation Center, Fort Rucker, AL.

Test Apparatus

A model 620 Visual Tracking Analyzer manufactured by Bolt, Beranek and Newman, Cambridge, MA, was used to administer the visual tracking test. The model 620 is capable of testing in either fixed or critical tracking mode but this investigation was limited to critical tracking data. The light spot is displayed on a horizontal unit 20 by 7.5 by 10 cm which contains a horizontal line of 64 light emitting diodes, each spaced 2.54 mm apart. The display unit is connected to a master control unit by a 15 foot wire cable with connectors at each end. The master control unit provides basic electronic circuitry, power supply, and the tester's unit. The tester's unit provides controls to (a) select the mode of tracking operation, (b) set the number of trials per testing block, (c) start a block of test trials, (d) enable the start of each test trial, (e) reject any unsuitable trial performance, and (f) conduct a standard system checkout to verify each of the system functions and displays and provide demonstrations of key features to each subject. Displays on the tester's unit provide status information about the state of the system, number of the current trial in a block, and the score for both the most recently completed trial and the current block average.

The subject controls the location of the light spot with lateral movements of a spring-loaded, finger operated joy-stick. One degree of stick deflection corresponds to a movement of 2.36 mm on the visual display. The control stick is mounted on a metal box 11.2 by 17.5 by 5 cm and it is connected to the visual display unit by a 6 foot wire cable with connectors at each end. The subject's control unit also contains a calibration thumb wheel and two trial start buttons, one button on either side of the control stick.

To measure the effective time delay, the test apparatus is operated in the critical tracking mode. The value of the system time constant at the end of a trial is the index of the subject's effective time delay for that trial. At the start of a trial the system automatically set the time constant at 500 milliseconds (ms). As the trial progresses the time constant is reduced at the rate of 10 ms per second until the light spot has deviated 2.5 cm from the center of the display and at the rate of 2.5 ms per second after the light spot has exceeded the 2.5 cm limit. As the size of the time constant decreases, the rate of movement on the display increases until the subject is unable to maintain the light spot location within the limits of the display. When the light spot location exceeds the limits of the display, the system stops the trial, displays the trial score and the current value of the block mean effective time delay on the tester's display, and signals an end of trial on the tester's status display. The tester must then record the trial score if it is desired and enable a new trial. The system is designed so that an attempt to enable a new trial at the end of a block will result in an end of block signal on the tester's status display.

Procedure

Subjects reported to a standard testing location according to a prescribed week long testing schedule. This testing schedule was worked out to provide continuity of testing over a five day period and to minimize the test activity interference with routine training. The second day of testing was used to give 40 trials of the critical tracking test in 4 blocks of 10 trials. The tester set the system to the system checkout/demonstration mode. When the subject reported for testing, he/she was seated at a table with the finger operated control stick. The tester then read through the following instructions:

> In this test your job is to control the movements of this light spot [tester points to light spot on visual display] with the control stick in front of you. Take hold of the stick in a comfortable position and move it right and left. Notice that the control moves the light spot back and forth on the display. Later, when you start the test, the light spot will move randomly right or left on the display from time to time. As a test progresses, the time between these random movements gets shorter and shorter and it gets harder and harder to control the position of the light spot. Finally, the light spot goes out of control, off the end of the display, and the system will freeze the light spot at the end of the display. Your score will be the time between the random movements when the light spot is frozen.
>
> (Tester note: Set the system in CRITICAL MODE.)
>
> Notice that the light spot is now frozen at the end of the display. Move the control stick and notice that the light spot does not move. When this happens that means the end of the test and I will read your score to you. To start a test you will find two buttons next to the control stick marked "START". After I say "Ready" you may push either button to start the test. When you release the button, the light spot will automatically move to the center of the display and the test will start. (Tester demonstrates.) Do you have any questions?
>
> You will repeat the test 40 times in the next hour. After each trial I will read your score to you. The smaller your score the better your performance. Your objective should be to get the smallest possible score in the fewest trials. To get a small score it is very important to keep the light spot as near the center of the display as possible. Do you have any questions on scoring?

Each trial was followed by 15 seconds rest and there was a 2 minute rest period after each block of 10 trials. At the end of each trial, the tester recorded the trial score, reported it orally to the subject, timed the rest interval, enabled the system for the next trial or block, and at the end of the rest time, announced "Ready" to signal the subject to start the next trial.

The subject participated in fixed difficulty tracking on the third and fourth days of testing before receiving a final test in critical difficulty tracking. On the third day the subject performed fixed difficulty tracking to establish levels of skill for the time-sharing test given on the fourth day. The time-sharing tests, lasting about 30 minutes, consisted of 45 trials of fixed difficulty tracking in 3 blocks of 15 trials, 1 block for each of 3 levels of tracking test difficulty. Following the time-sharing tests each subject received 5 trials in critical tracking mode as a final test of tracking skill.

Data

The tester recorded the effective time delay score for each of the 40 initial and the 5 final trials of critical tracking. Recorded on a standard form specifically designed for use with critical tracking in the aviator selection research program, the critical tracking scores were later transcribed to standard 80 column computer card image forms, checked by a second person, and keypunched with verification. A special FORTRAN program was prepared to compute means and standard deviations for the 9 blocks of 5 trials and to compute the z-score conversion of the MSSD measure from all the data in the first 40 trials.

Design

A two-way categorization was used as the design of the subsequent analyses. The two categories were type of subject, trainee versus attritee, and type of performance, random (z-score less than 1.96) and nonrandom (z-score greater than or equal to 1.96); nonrandom in this case means that the data contained a linear or higher order polynomial trend.

Data Analysis

The first step in the data analysis was to compute a chi-square to test the hypothesis that frequency of classification of type of performance was not dependent on student category. Acceptance of this null hypothesis of no dependency would be used as evidence for employing a least squares analysis of variance procedure with the observed cell frequencies as the best estimates of the proportions in the population. Rejection of the null hypothesis of frequency of classification would indicate a need to employ methods to adjust the degrees of freedom in the analysis of variance procedures.

A 2 between-, 1 within-subjects repeated measures analysis of variance was used to test hypotheses about the equality of (a) mean effective time delay and (b) the standard deviation of effective time delay for the five trial blocks. Any effect in the chi-square test or the analyses of variance was considered statistically significant at the conventional .05 level.

RESULTS

The z-score transform from each subject's data was used to classify his/her performance as random or nonrandom. If the z-score was less than 1.96 the performance was classified as random. Any performance with a z-score greater than or equal to 1.96 was considered nonrandom, i.e., the data contained a trend. As a one-tailed test, this rule would result in a Type I classification error about 2.5% of the time. Table 2 gives the breakdown of number of subjects in each cell of the two by two student category by performance type matrix.

Table 2

Breakdown of Number of Subjects

Student Category	Type of Performance Random	 Nonrandom	Total
Trainee	11	9	20
Attritee	2	7	9
Total	13	16	29

A chi-square analysis was used to determine if the classification of random versus nonrandom performance was dependent on student category. The marginal totals were used to define the expected cell values because there was no prior reason to expect a particular breakdown pattern. The results of the chi-square analysis revealed no statistically significant dependency in the observed breakdown of number of subjects ($\chi^2 = 1.53$, $p > .10$, 1 df). This result was interpreted as evidence for using a least squares analysis of variance for unequal cell frequencies with mean effective time delay (Winer, 1971).

Mean Effective Time Delay

The measure of skill in the critical tracking test was effective time delay. Means for each subject for nine blocks of five trials in the 40 practice and 5 final test trials were analyzed with analysis of variance (Table 3). The hypothesis of an interaction between student category and type of performance across the nine blocks of five trials was not confirmed by the mean effective time delay measure. The analysis of variance revealed statistically significant main effect for blocks of trials which indicated that average performance had improved with practice (Figure 3).

There were two statistically significant interactions for the mean effective time delay. Student category interacted with blocks of trials

1202

Table 3

Analysis of Variance Summary for Block Means of Effective Time Delay

Source	df	Mean Square	F-Ratio	$\hat{\omega}^2$
Total	260	2696.22		
Between Subjects	28	12648.45		
Student Category (A)	1	5024.53	.43	-
Performance Type (B)	1	35276.90	2.96	.004
A x B	1	16242.16	1.36	.001
Error	25	11504.52		
Within Subjects	232	1495.09		
Blocks (C)	8	24921.2.	43.38***	.470
A x C	8	1513.15	2.63*	.023
B x C	8	2314.58	4.03**	.043
A x B x C	8	245.53	.43	-
Error	200	574.51		

*$p < .025$

**$p < .01$

***$p < .001$

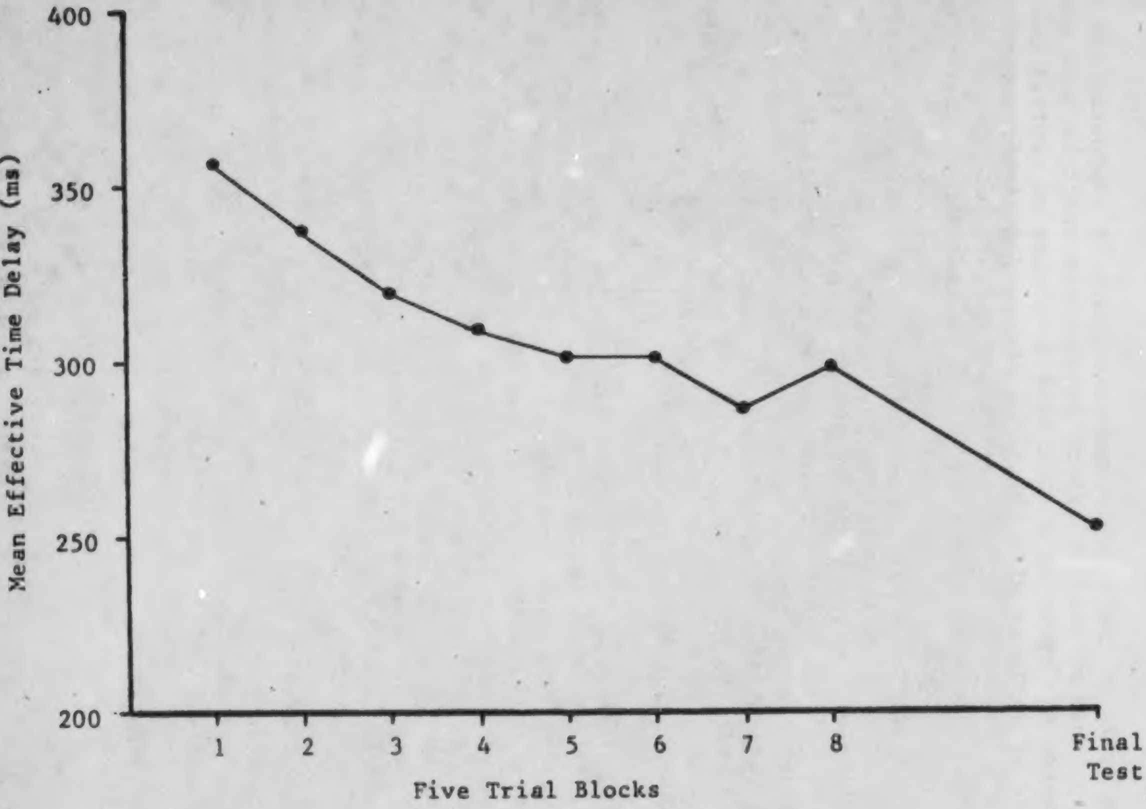


Figure 3: Improvement of mean effective time delay as a function of blocks of trials.

(F = 2.63, p < .025, df = 8, 200, proportion of variance = .023). As shown in Figure 4, the source of this interaction effect was the larger effective time delay means for the attritees on the first three blocks of trials. The other statistically significant interaction was type of performance with blocks of trials (F = 4.03, p < .001, df = 8,200, proportion of variance = .043). Figure 5 shows that the source of this effect was the difference in slopes between the two types of performance which indicates the greater rate of learning or degree of effort for the nonrandom group.

Variability of Performance

Analysis of the block standard deviations supported the hypothesis that a measure of variability of performance would be more sensitive to differences of degree of effort than a measure of central tendency. Analysis of variance with the block standard deviations confirmed the hypothesis that student category would interact with type of performance across the blocks of trials and also revealed other significant differences (Table 4). Figure 6 shows mean standard deviation as a function of student category and type of performance across blocks of trials. One striking feature of these plots is the extreme differences in block to block variability of the two attritee groups in relation to the variability of the trainees. The random trainee group exhibits the least block to block variability and the nonrandom trainee group gives strong evidence of improvement of variability with practice. Finally, the equivalence of the mean standard eviation on the final test for each of the four groups strongly suggests that factors other than differences in level of tracking skill are influencing the performances of the members of the different groups.

Some caution must be used in interpreting the variability of the random attritee group because the group has only two subjects. However, these two subjects also have the greatest total variances of any of the subjects in the design matrix (Table 5). As would be expected from an insection of the group plots in Figure 6, the size of the total variances in Table 5 is correlated with group membership. This correlation is supported by the significance of the between subjects effects in the analysis of variance summary (Table 4). Table 6 gives the mean standard deviations for each of the main effects and the interaction in the two by two student category by type of performance part of the design. Finally, Figure 7 shows the interaction of type of performance across blocks of trials on mean block standard deviation. The interesting feature of this interaction is the increasing variability trend of the random versus the decreasing variability trend of the nonrandom groups. This difference of trend of variability as a function of type of performance is strong support for the hypothesis that MSSD is an indicator of differences in performance patterns.

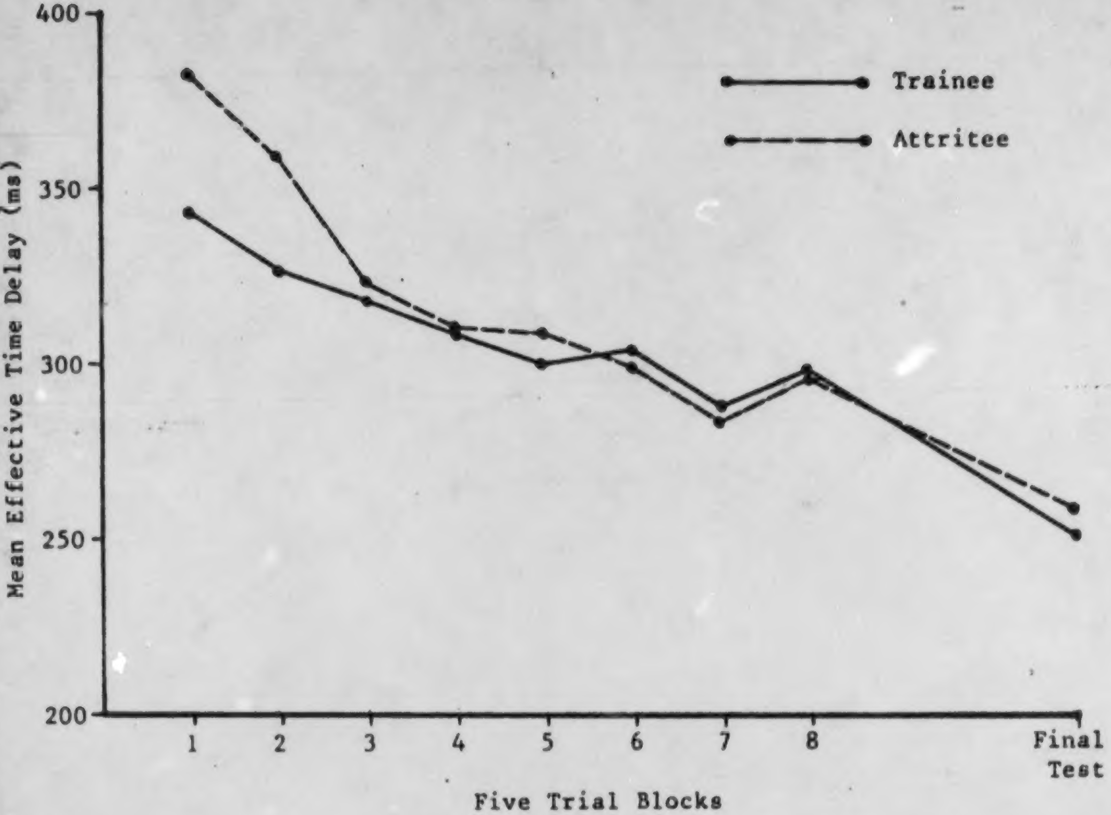


Figure 4: Interaction of Student Category on Mean Effective Time Delay Across Blocks of Five Trials

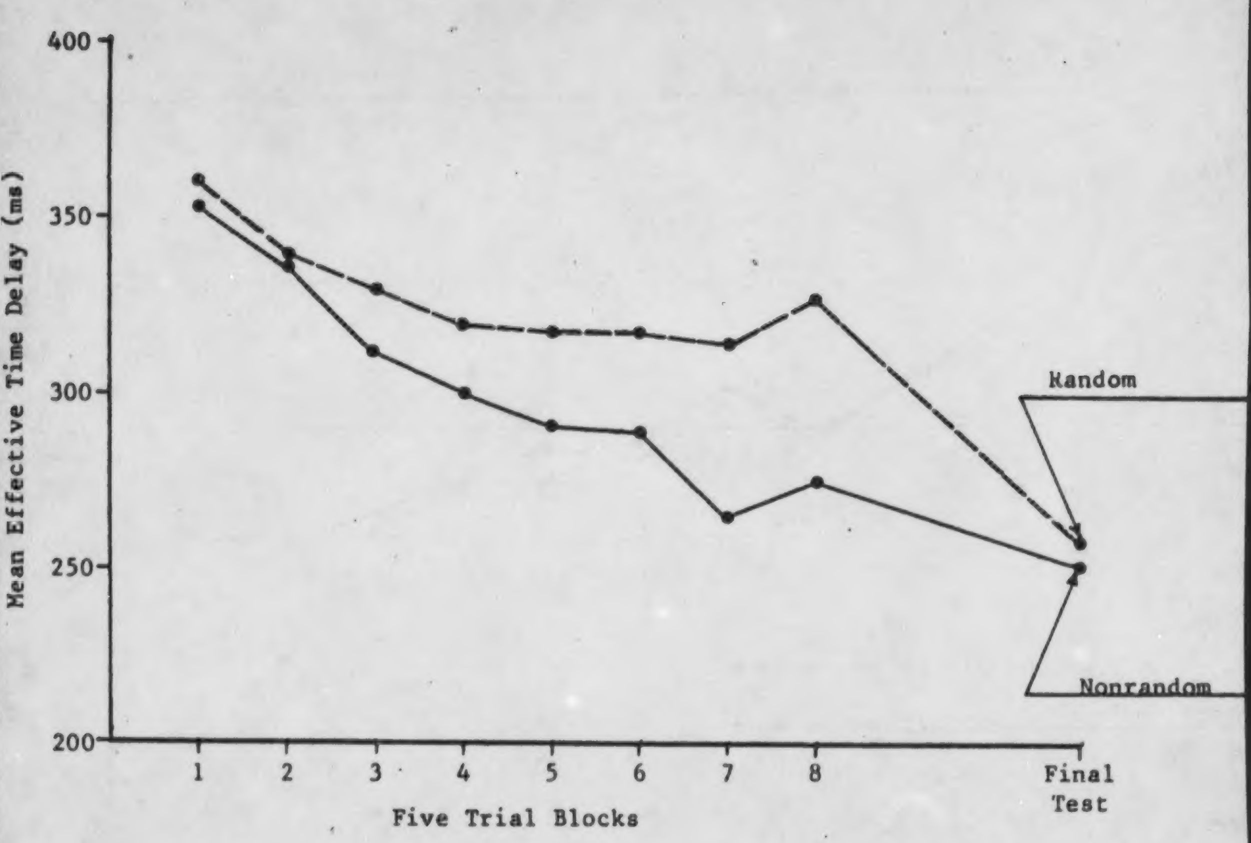


Figure 5: Interaction of type of performance across blocks of trials on mean effective time delay.

Table 4

Analysis of Variance Summary for Block Standard Deviations of Effective Time Delay

Source	df	Mean Square	F-Ratio	$\hat{\omega}^2$
Total	260	308.54		
Between Subjects	28	420.86		
Student Category (A)	1	1734.88	6.53**	.018
Performance Type (B)	1	1193.61	4.49*	.011
A x B	1	2213.94	8.33***	.024
Error	25	265.67		
Within Subjects	232	294.67		
Blocks (C)	8	385.25	1.48	.012
A x C	8	201.35	.77	--
B x C	8	524.14	2.01*	.026
A x B x C	8	935.17	3.59***	.067
Error	200	260.34		

*p <.05

**p <.025

***p <.001

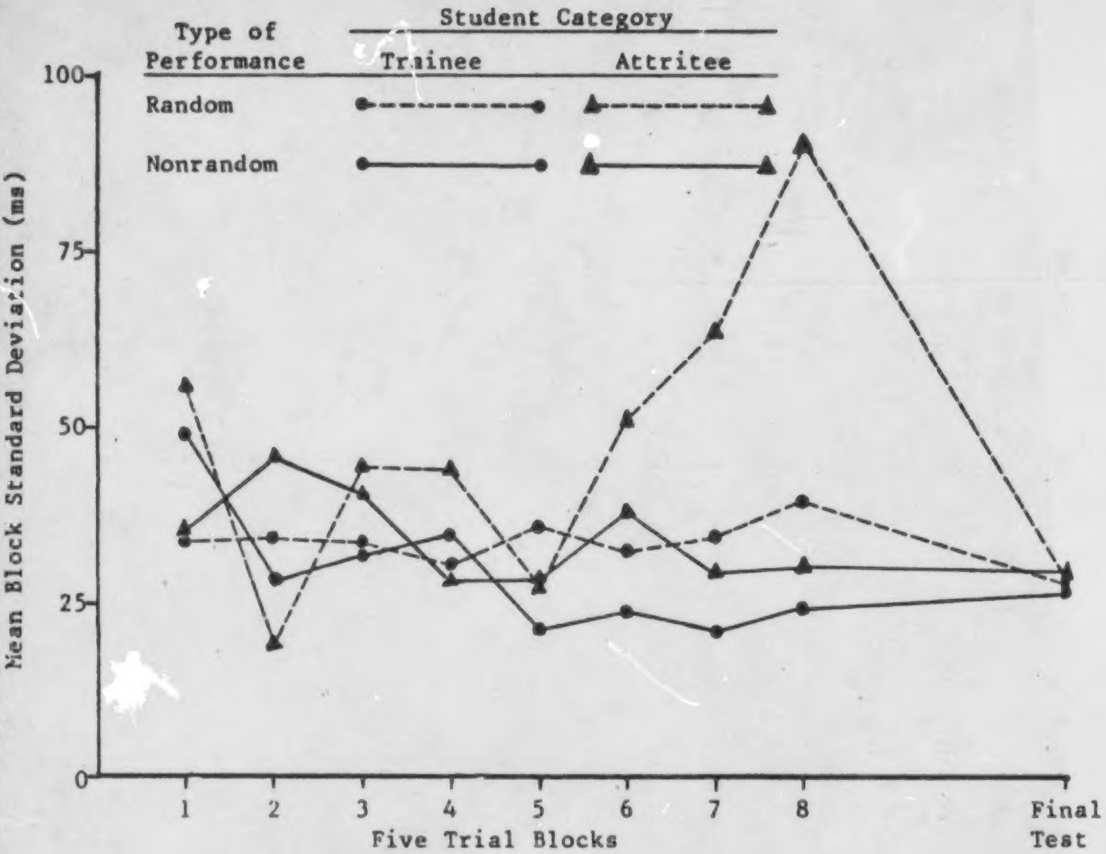


Figure 6: Interaction of student category with type of performance across blocks of trials on mean block standard deviation.

Table 5

Sum of Block Variances for each Subject

	Type of Performance	
Student Category	Random	Nonrandom
Trainee	6509.8	4890.5
	6850.4	5519.7
	7570.0	6520.2
	10040.3	8010.5
	10903.3	8649.5
	11549.0	9439.6
	12211.4	12700.4
	14599.8	15989.9
	17152.5	18569.7
	18239.4	
	19180.8	
Attritee	21580.9	8309.4
	29657.8	9499.5
		11220.6
		13429.4
		14869.6
		16350.1
		18228.8

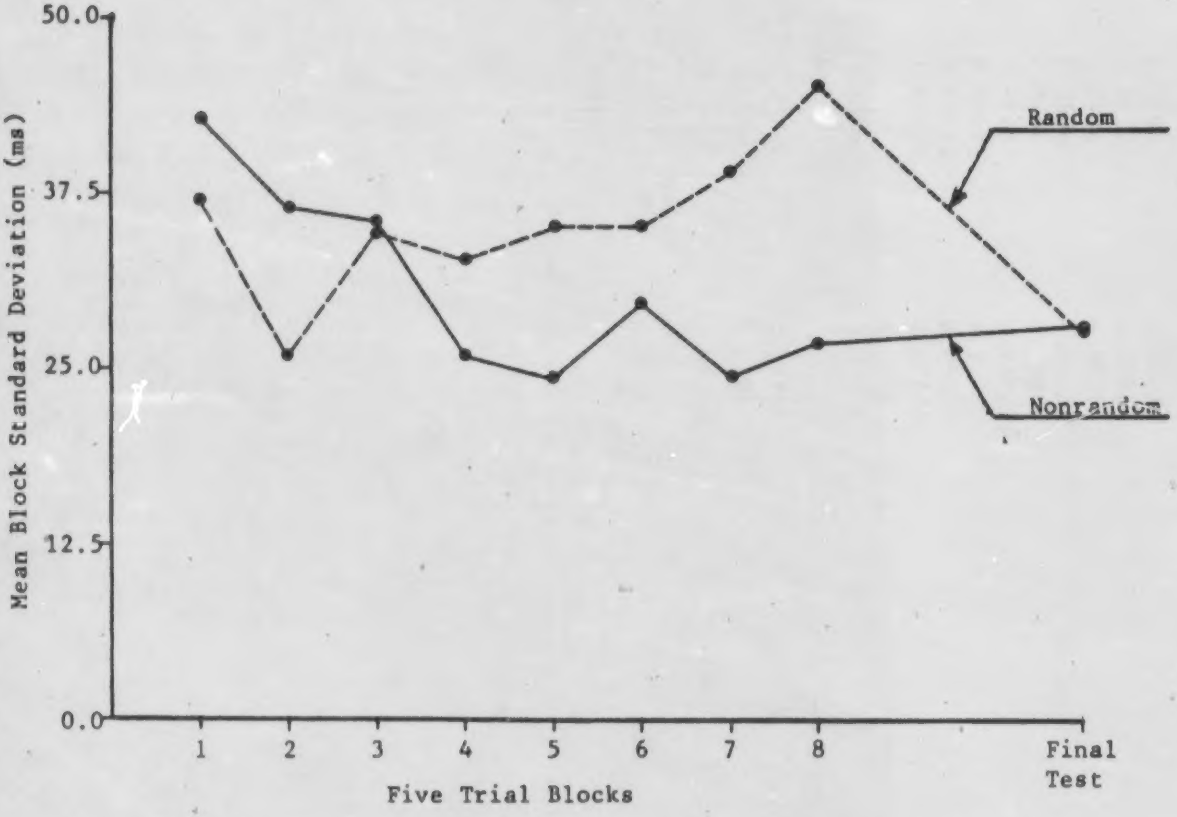


Figure 7: Interaction of type of performance across blocks of trials on mean block standard deviation.

Table 6

Mean Block Standard Deviations for Student Category by Type of Performance

	Type of Performance		
Student Category	Random	Nonrandom	Combined
Trainee	33.39	29.16	31.49
Attritee	47.68	34.03	37.06
Combined	35.59	31.29	33.22

DISCUSSION

The main hypothesis of this investigation was that degree of effort would be a source of confounding in tracking test performance. The results confirmed this hypothesis if degree of effort varies with motivation to perform and differences in motivation depend on student category. The major source of this confounding was differences in variability of performance as a function of number of test trials. The source of the interaction is most clearly apparent in the comparison of type of performance with student category across the blocks of trials on mean block standard deviation (Figure 6). Inspection of mean effective time delay interactions shows that mean effective time delay is correlated with variability of performance which is consistent with Poulton's hypothesis of a shift in performance strategy.

A second hypothesis was that inadequate practice was a source of confounding in the measurement of level of achievement in previous research. In this investigation, level of achievement is represented by mean effective time delay and Figure 3 clearly shows a large improvement of this measure, even after the eighth trial block. A comparison of mean effective time delay from this investigation and a previous study by Pew et al. (1977) with the same tracking test further supports the hypothesis of inadequate practice. In the Pew et al. study 92 students in Air Force Undergraduate Pilot Training at Williams Air Force Base, AZ, performed 10 trials of the critical tracking test. Table 7 is a comparison of the mean and standard deviation of effective time delay for the last 7 trials of the Pew et al. study with the means and standard deviations of 5 trials blocks and the final test for trainees in the present study. (Trainees were used for comparability of populations.) The important comparisons in Table 7 show that there were no significant differences between the Pew et al. results and those of this investigation on the first 4 blocks of trials.

Table 7

Means and Standard Deviations of Effective Time Delay from Pew et al. (1977) and Blocks of Trials for Trainees from the Present Investigation

Measure	Pew et al.[a]	Block[b]								Final Test
		1	2	3	4	5	6	7	8	
Mean	340.3	344.0	327.6	318.5	308.5	300.0*	303.1*	288.7*	298.9*	250.9**
S.D.	52.3	66.3	59.4	59.4	58.2	55.3	54.1	58.4	60.1	40.0

$^{a}\underline{n} = 92$; $^{b}\underline{n} = 20$; $^{*}\underline{p} < .05$; $^{**}\underline{p} < .001$

The results of this investigation indicate that mean square successive differences (MSSD) should be a useful statistical tool in subsequent research. Although MSSD was employed in the present investigation as a one-tailed test to indicate polynomial trends, significant negative values of the *z*-score derived from MSSD would indicate that the data contained systematic cyclic or periodic trends, i.e., trends describable with trigonometric functions. This latter feature makes MSSD especially useful in the analysis of tracking performance from continuous control tasks where periodic features of the data may indicate important differences in operator control behaviors. With a significant positive or negative *z*-score from the MSSD measure, the data analyst is justified in a detailed search for the sources of the specific polynomial or periodic trends in an individual set of data.

Research is needed to establish the predictive validity of differences in patterns of performance from the tracking test for overall success in pilot training. Interviews with instructor pilots have indicated that lack of motivation is frequently a source of inadequate student progress in Army helicopter pilot training. This instructor pilot observation is supported by two sources of additional evidence. First, some 50% of all attrition in the Army helicopter pilot training program results from resignations (Elliot & Joyce, 1978). Furthermore, motivation was identified as a major factor among resigning students. Second, an unreported exploratory investigation at the US Army Aviation Center found a correlation of .78 between instructor pilot ratings of basic student pilot qualities, e.g., motivation, judgment and the like, on daily grade sheets from early primary training and subsequent eliminations from advanced training. This evidence suggests that the present approach may yield a substantial reduction in the residual variance of the aviator trainee selection testing process.

The approach used in this investigation also presents some interesting possibilities for further research in aviator trainee selection and management methods. For example, detailed analyses of individual performance trends were not accomplished in the present investigation. However, the logical analysis of degree of effort depicted in Figure 1 indicates that differences in such trends should further differentiate among types of performance and the associated performance strategies. One interesting hypothesis is that learning behavior, i.e., performance strategy, in a simple tracking test would predict learning behavior in more complicated tasks, i.e., performance in aircraft control.

Cronbach and Snow (1977) evaluate the hypothesis of prediction from learning behavior in these terms:

> If individual differences prove to be stable and predictable, one can capitalize on findings from the experiment in which learning is observed only for a short time, perhaps on just one task or topic. If individual differences are radically altered during learning... the short-term experiments...will not give practically useful conclusions. Under this hypothesis, persons who learn most efficiently, among a group all of whom have become familiar with the problem, would not generally be the ones who learned most efficiently at the outset; hence, they would not have been among the most successful learners in a short experiment (p. 126).

The major issue is whether attitudinal differences such as motivation which are reflected in the degree of effort measurement procedures are relatively stable characteristics of an individual's learning behavior. As a test of the Cronbach et al. hypothesis in a subsequent investigation, the methods of this investigation will be employed to predict performances of these same subjects in fixed stability tracking, time-sharing, and a job-sample test administered on the UH-1 flight simulator.

REFERENCES

Brownlee, K. A. Statistical theory and methodology in science and engineerin (2nd Ed.). New York: Wiley, 1965.

Cronbach, L. J. and Snow, R. E. Aptitude and instructional methods: a handbook for research in interactions. New York: Wiley, 1977.

Damos, D. L. Development and transfer of timesharing skills (ARL-77-19/AFOSR-78-134). Urbana-Champaign: Aviation Research Laboratory, Institute of Aviation, University of Illinois, July 1977. (DDC No. AD-A-050-255)

Gopher, D. and North, R. A. The measurement of operator capacity by manipulation of dual-task demands (ARL-74-21/AFOSR-74-15). Urbana-Champaign: Aviation Research Laboratory, Institute of Aviation, University of Illinois, October 1974.

Hart, B. I. Significance levels for the ratio of mean square successive difference to the variance. Annals of mathematical statistics, 1942, 13, 445-447.

North, R. A. Task components and demands as factors in dual-task performance (ARL-77-2/AFOSR-TR-77-0519). Urbana-Champaign, IL: Aviation Laboratory, Institute of Aviation, University of Illinois, January 1977. (DDC No. ADA 038634).

North, R. A., Harris, S. D. and Owens, J. M. Test-retest reliability of individual differences in dual-task performance (NAMRL-1248). Naval Air Station, Pensacola, FL: Naval Aerospace Medical Research Laboratory, July, 1978.

Pew, R. W., Rollins, A. M., Adams, M. J., and Gray, T. H. Development of a test battery for selection of subjects for ASPT experiments. (Technical Report No. BBN 3585). Cambridge, MA: Bolt, Beranek, and Newman, November 1977.

Poulton, E. C. Tracking. In Bilodeau, B. A. and Bilodeau, I. McD. Principles of Skill Acquisition. New York: Academic Press, 1969, 287-318.

Rose, A. M. Human infomation processing: An assessment and research battery (Technical Report No. 46). Ann Arbor, MI: Human Performance Center, University of Michigan, January 1974.

Winer, E. J. Statistical principles in experimental design (2nd Ed.). New York: McGraw-Hill, 1971.

Table 1

Editiorial Changes to Preprint Draft of the Revised FAST

Draft page	Item	Comment
MANUAL		
5	Table 1	Under heading "FAST Booklet..." the phrase "Stick and Rudder" should be replaced with the word "Cyclic"
12	Directions	The paragraph beginning "THE SPECIAL DIRECTIONS..." should be changed as follows: Strike all the wrods (1)/In the third line beginning at "...A CORRECTION " and ending in the sixth line at the end of the sentence ending with the words "...CORRECTION FACTOR IS APPLIED." (2) Insert the folling in place of the words delete in (1) above: insert-- Following"...OR IF"/ THE TEST SCORE WILL BE ADJUSTED BY SUBTRACTING A FRACTION OF THE WRONG ANSWERS FROM THE NUMBER OF RIGHT ANSWERS. EVEN ON THOSE TESTS WHERE THE SCORE IS ADJUSTED FOR WRONG ANSWERS."YOU SHOULD..."
TEST BOOKLET		
2	Instructions	Insert following the last parapgraph: DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

SECTION 17

SIMULATORS AND SIMULATION

EVALUATION OF TROUBLESHOOTING SIMULATOR

Dale A. Steffen and Anita S. West
Denver Research Institute
University of Denver
Denver, Colorado 80210

INTRODUCTION

For several years the Air Force has been involved in the investigation of and experimentation with simulators for teaching hands-on maintenance tasks [Miller, 1976]. Investigations have shown that simulators provide troubleshooting instruction which is at least equal to that afforded by actual equipment while offering additional opportunities such as increased individualized instruction by enabling more practice with job skills, increased assistance early in the instructional process via CAI techniques, increased consistency in student evaluations, and decreased equipment costs associated with breakage, obsolescence, and the need for special purpose training equipment. Furthermore, the application of computer-assisted performance training to troubleshooting instruction provides realistic feedback while manipulating the real time variables, for example, an induced malfunction from a student action that might normally occur only after several hours of time in actual equipment.

Due to the obvious cost-effectiveness, most computer-controlled simulators employing CAI and CMI have been constructed to replace sophisticated, costly equipment requiring high level skill training. On the other hand, simulators which teach fundamental principles and provide training of basic skills are generally not controlled by computer or, if controlled by computer, exhibit little or no CAI and CMI instructional techniques. If CAI and CMI techniques are to be utilized for such applications, it is necessary from a cost standpoint that the simulator exhibit general purpose properties that allow interchangeable simulation modules on a mainframe console containing the computer to provide usage in a wide variety of job skills training.

TRAINING CARREL

To continue the investigation of the utility of general purpose simulation in formal technical training environments, the Air Force, through coordinated efforts between personnel of the Human Resources Laboratory (at Lowry Air Force Base) and the Denver Research Institute developed a computer-assisted performance training carrel. The carrel contains interface circuitry to/from a PDP-11 computer (a minicomputer manufactured by Digital Equipment Corporation) and a PLATO IV terminal (the University of Illinois plasma screen terminal) designed with I/O bus circuitry for control of peripherals (various devices such as switch closures, digital-to-analog converters, waveform generators, etc.).

This instructional carrel has a random access slide projector to operate in conjunction with the PLATO screen which also contains a Touch-Panel capability. Additionally, the carrel contains provisions for "plug-in" simulation panels which offer flexibility in implementing a variety of simulations via several panels containing different peripherals all of which can be controlled by the I/O circuitry for simulating probes, displays and switches, as well as voltage, resistance and current parameters for equipment familiarization, instruction, troubleshooting and maintenance.

The carrel was designed in a way that allows experiments involving the formation of concepts and the retention of learning to be enhanced by capabilities formed by a variety of visual and auditory cues and stimuli. The communication between the computer devices permits control to be passed to the more efficient of the two computers for each specific application.

TRAINING CARREL AS A TROUBLESHOOTING SIMULATOR

An experiment was designed to evaluate both the performance training carrel and each of its component parts -- software and hardware. The approach for the evaluation was to prepare a short course demonstration of an existing instructional module in order to compare carrel instruction with traditional instruction. The instructional module presented was a "Troubleshooting Fundamentals" module in an Air Force Air Training Command Electronics Principles course. As such, the "plug-in" panel implemented for use in this module resulted in the carrel serving as a troubleshooting simulator.

The selected module allowed a study of each of the functions of the carrel with a minimum of disruption in the course where it appeared. The technical level allowed a sophistication of simulated test equipment and training circuitry that required the utilization of both analog and digital controllers/sensors in conjunction with the process control/sampling minicomputer (PDP-11) to interpret data to the instructional computer (PLATO IV system).

A group of forty students was selected, twenty of whom were selected at random to receive traditional lectures/demonstrations of the selected course module. The remaining twenty students received course delivery via the troubleshooting simulator. Course objectives and instruction were paralleled in both types of delivery to allow the use of the same criterion measurement on both groups. Thus, the emphasis was placed on the evaluation of hardware and software functions of the simulator, and no attempt was made to optimize the instructional strategies for delivery on the simulator.

The nature of the module chosen required the simulation of a PSM-6 multimeter and various DC trainer "schematic boards" which could easily be interchanged on the simulator panel. The schematic boards exhibited essentially the same appearance as those used in a traditional delivery of the course. Thus, the use of the boards were familiar to students. Figure 1 illustrates the interaction between the student and the performance carrel during the delivery of the instructional module.

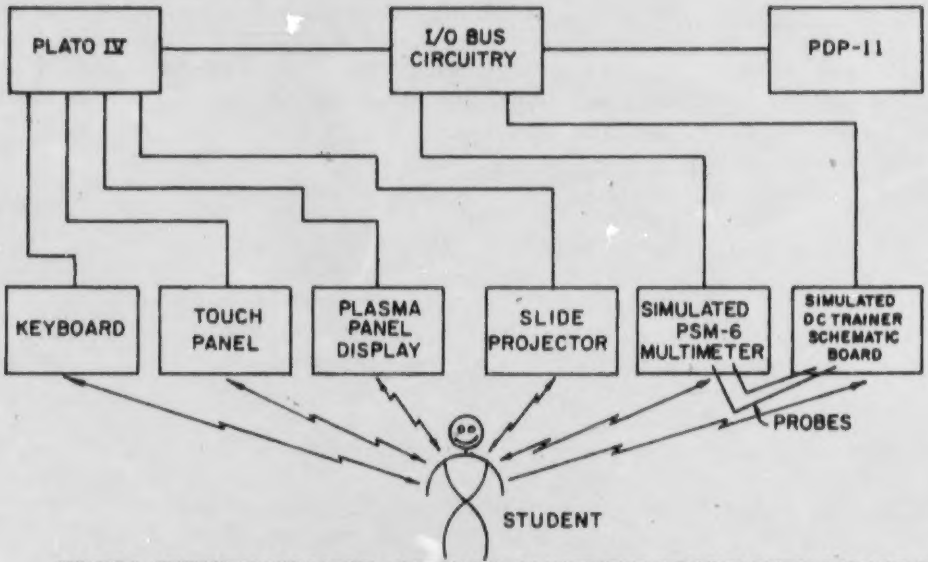


FIGURE 1. PERFORMANCE CARREL SYSTEM DIA. ILLUSTRATING STUDENT INTERACTION DEVICES

The performance training was presented to the student in two modes of instruction, namely, the presentation of theory involving troubleshooting fundamentals via programmed instruction, and secondly, a mode whereby the student was given a troubleshooting problem and had to proceed by interacting with the simulated equipment independent of programmed instruction. In the first mode, the student had to be responsive to computer actions. In the second mode, the computer had to be responsive to student actions.

The computer monitored two classes of error made by the student during delivery of the module. The first class included misuse of the equipment, while the second class dealt with incorrect troubleshooting logic on the part of the student. For each class of error, the PLATO IV system responded with remedial instruction. Several schematic boards representing various D.C. trainer circuits of different levels of difficulty were presented to the student in a manner to provide the opportunity for successfully completing a problem on one level of difficulty before advancing to a more difficult level.

Since only the PLATO IV has control over these devices in the carrel which communicate directly with the student, the software was developed so that the PLATO IV system controlled the course delivery in a way that the interaction of the student was with the PLATO IV terminal only. The PLATO IV touch panel, slide projector, and keyboard were the only peripherals used during this mode of instruction.

For the second mode of instruction, when the computer had to be responsive to the student, the function of the PLATO IV system was to establish which schematic board(s) would be used and which malfunction(s) were to be implemented. Also, the PLATO IV presented all remedial instruction required during this phase of performance training. In this case, the PDP-11 was used for the simulation of devices on the carrel control panel and for the time history monitoring of student actions.

The software requirements established for the PDP-11 allow it to fulfill three major functions:

a. To simulate equipment (that is, the multimeter and DC trainer schematic boards) based on tables which define responses to student actions at the control panel of the performance carrel.

b. To monitor the actions of the student.

c. To communicate student actions or status changes to the training computer and accept instructions from the training computer.

Figure 2 shows a typical circuit which is simulated on a schematic board. The solid line circles are accessible test points provided in this schematic board for probe insertion. The dashed line circles are additional test points in the control panel which are available to allow flexibility in circuit lay-out of other schematic boards. There are 28 possible test points for use in lay-out. The lamps and switch are shown only schematically on the schematic board. However, actual lamps and a switch are mounted directly below the schematic board. It was decided to mount these components off the actual circuit diagram to allow maximum versatility in the location of these simulated components on various schematic boards. Schematic board I.D. switches are mounted on the control panel. Varied combinations of holes can be drilled in the schematic boards which determine which of the I.D. switches are activated. This allows the identification of the schematic board on the control panel.

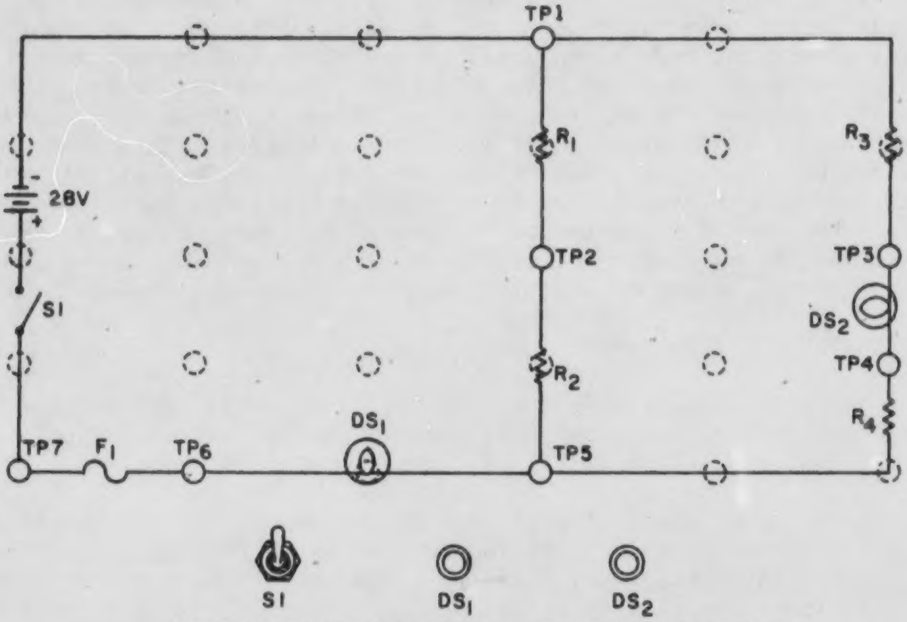


FIGURE 2. EXAMPLE OF TYPICAL SCHEMATIC BOARD CIRCUIT

RESULTS OF EVALUATION

The results of the analysis of achievement for the experimental group and a control group using paper and pencil programmed text with a circuit "trainer" showed no significant differences. The baseline measures of math scores, reading level and Block 1 test scores for each group were not statistically different. The learning and performance measures were also equivalent for both groups, e.g., the Block 2 test scores, the retention test scores, the practical performance test scores, and the time taken to complete the module were not significantly different.

CONCLUSIONS

The equivalent achievement measures and positive attitudes of the students indicate a potentially very satisfactory teaching mechanism. The experimental system has provided insight into many of the contingencies of hardware/software interfacing and instruction in troubleshooting by a performance training simulator. Transfer of this information to the design and implementation of additional systems, and more complete utilization of the information generated by the CAI capability of the carrel will provide the Air Force and the training community in general with a powerful instructional tool.

In order to retain the integrity of the experimental design for measuring the effectiveness of carrel versus programmed instruction, every effort was made to hold all of the other independent variables constant. As a result of this constraint, the carrel instruction closely followed the existing study guide and lesson plan. However, the plans of instruction for delivery by one method of instruction may not be optimal for another mode of delivery. That is, alternative instructional strategies may be more appropriate for carrel presentation. Methods of presentation for carrel delivery should be examined, since the CAI-simulator delivery mechanism may be enhanced by a different format, order of material, or type of presentation.

REFERENCES AND BIBLIOGRAPHY

Gray, G.C. and Steffen, D.A., "PLATO IV/PDP-11 I/O Bus Line Control System", DRI Report #2640, University of Denver, Denver Research Institute, Denver, Colorado, February 1974.

Kargo, D.W. and Steffen, D.A., "Performance Training Carrel for Electronics Principles Course", AFHRL-TR-76-62(I), Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, September 1976.

Miller, Gary G., "A Survey of the Potential Application of Generalized Simulation to Air Force Technical Training," AFHRL/TT, Lowry Air Force Base, (1976).

Wasmundt, K.C. and Steffen, D.A., "Software for Performance Training Carrel," AFHRL-TR-76-62 (II), Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, September 1976.

West, A.S., Wasmundt, K.C., Lantz, A.E., Steffen, D.A. and Miller, G.G., "Performance Training Carrel Evaluation", AFHRL-TR-76-62 (III), Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, (1976).

METHODOLOGY FOR EVALUATING OPERATOR PERFORMANCE ON TACTICAL OPERATIONAL SIMULATOR/TRAINERS

Charles W. Howard, Ph.D
United States Army Research Institute
for the
Behavioral and Social Sciences - Ft. Bliss, Texas

October 1978

Paper Presented at the 20th Annual Conference
of the Military Testing Association

METHODOLOGY FOR EVALUATING OPERATOR PERFORMANCE ON TACTICAL OPERATIONAL[1] SIMULATOR/TRAINERS

BY

Charles W. Howard, Ph.D.
US Army Research Institute for the Behavioral and Social Sciences
Ft Bliss, Texas

ABSTRACT

This paper discusses a method presently being developed for evaluating operator performance on a Tactical Operational Simulator/ Trainer (TOS/T).

Quantifying operator performance has been an important area of research for several years. A common problem for many of these performance studies was their inability to duplicate standard scenarios or sets of events. Compounding the problem was their inability to obtain baseline measures of these scenarios while the system operated in an automatic mode.

Recent technological advances have made it possible, however, to duplicate standard scenarios, and to execute and reexecute them on computer-driven training simulators. These scenarios also can be executed in either an automatic mode, which is non-operator dependent, or a semi-automatic mode, which requires some operator performance in a more systematic fashion.

The automatic mode capability enables researchers to establish baseline measures of scenarios. These baseline measures then become the "yardstick" for evaluating operator performance in the semi-automatic mode. The presentation of standard scenarios or sets of events permits comparisons of operator performance for repeated trials, in addition to comparisons among operators. Furthermore, this method of evaluating operator performance enables developers of training scenarios to produce objective-oriented training programs.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

INTRODUCTION

Past research regarding operator performance on radar systems has focused on man's ability to detect, track and identify targets, and make engagement decisions. The research has been conducted in various environments; real and simulated, and computer assisted and non-computer assisted. None of these environments though have permitted researchers to capture the sequence of system actions and use these actions as a baseline for comparison with actual operators' actions. In an automatic mode, however, system software could be used to generate an "ideal" sequence of actions which could be compared to operators' actions. Operators' performance could be evaluated by noting the difference between the system's (model) actions and the operators' (actual) actions.

The Army's newest major air defense weapon system, PATRIOT, is "the first truly automated, fully software driven air defense weapon system!" Because the PATRIOT system software is truly automatic, it can be used to generate an "ideal" sequence of actions for research and training purposes.

The purpose of this paper is to discuss a method presently being developed to assess and evaluate operator performance on the PATRIOT system console. The research and training will be performed on a Tactical Operational Simulator/Training (TOS/T), (see Figure 1), which will permit a scenario to be run and re-run in automatic mode and will allow comparisons to be made between model and actual actions.

METHOD

Several tasks will have to be completed before the collection and evaluation of operators' performance data can be begun. These prerequisite tasks include the following:

1. Developing an Automated Training Scenario Generated Program (ATSGP). The ATSGP should minimize the time it will take to develop training scenarios.

2. Designing objective based scenarios. These scenarios will be used to train console operators to locate and use the console buttons, i.e., the operators will be trained to determine where the buttons are and how they elicit various system software actions.

3. Modifying existing system simulation software. The system simulator software, as it is now, will have to be modified before system data or operators' actions data can be collected.

4. Preparing training material. A set of training materials will be prepared to train console operator tasks.

5. Validating training materials. The training materials will be validated before actual collection of the data.

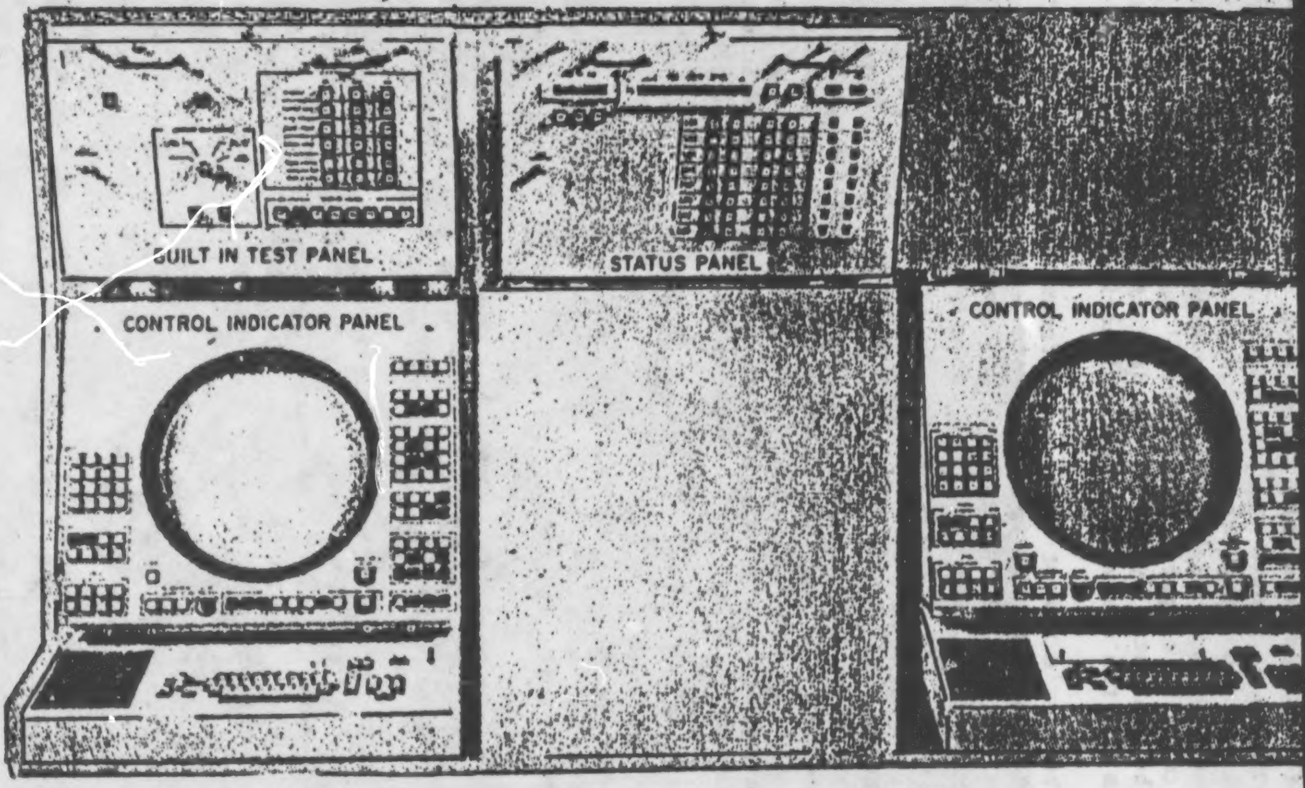

BUILT IN TEST PANEL
STATUS PANEL
CONTROL INDICATOR PANEL
CONTROL INDICATOR PANEL

Figure 1. Operator console in Patriot Weapon System

Once these prerequisite tasks have been completed the collection of data will begin. Data for the system will be collected in real time for standard scenarios. Data for the operators will be collected at preplanned intervals of one-tenth of a second for standard scenarios. Although the data will be collected in real time, it will be evaluated offline, because the additional requirements demanded by the evaluation procedure may degrade realtime simulation performance.

The data to be collected are discussed in the next section of this paper.

Proposed Model

The variables to be manipulated in the proposed model include: The number of threats (T_i) in a given scenario (see Figure 2), the number of resources spent per threat, and the effect of asset value.

The ratio of incapacitated threats (I/T_i) yields a measure of relative fire section effectiveness (EFF). Since each threat may attack assets, the EFF measure also accounts for threats that penetrate. Therefore the product of the ratio of incapacitated threats and the reciprocal of penetrators yields a refined EFF measure, depicted as:

$$EFF = \frac{I}{T_i} \cdot \frac{1}{P}$$

where $P > 0$.

If $P = 0$, however, P is reset to 1.

To simplify this reset problem, due to the possibility of a given threat not penetrating, the model is revised as shown.

$$EFF = \frac{I}{T_i} \cdot \frac{1}{TP}$$

where $TP = 1 + P$.

A given set of training scenario stimuli may result in a battle time of only twenty seconds. Battle time is defined as the difference between the time the first vehicle appears on the screen and the last vehicle moves off the screen. The EFF measure to be used in analyzing training scenarios requires a unit of time for system response to threat engagement (te) status and a unit of time to incapacitate (ti) the threat. This amount of time (ti) is incorporated into the model as the reciprocal of threat time ($1/t_i$). Therefore, the model takes the form;

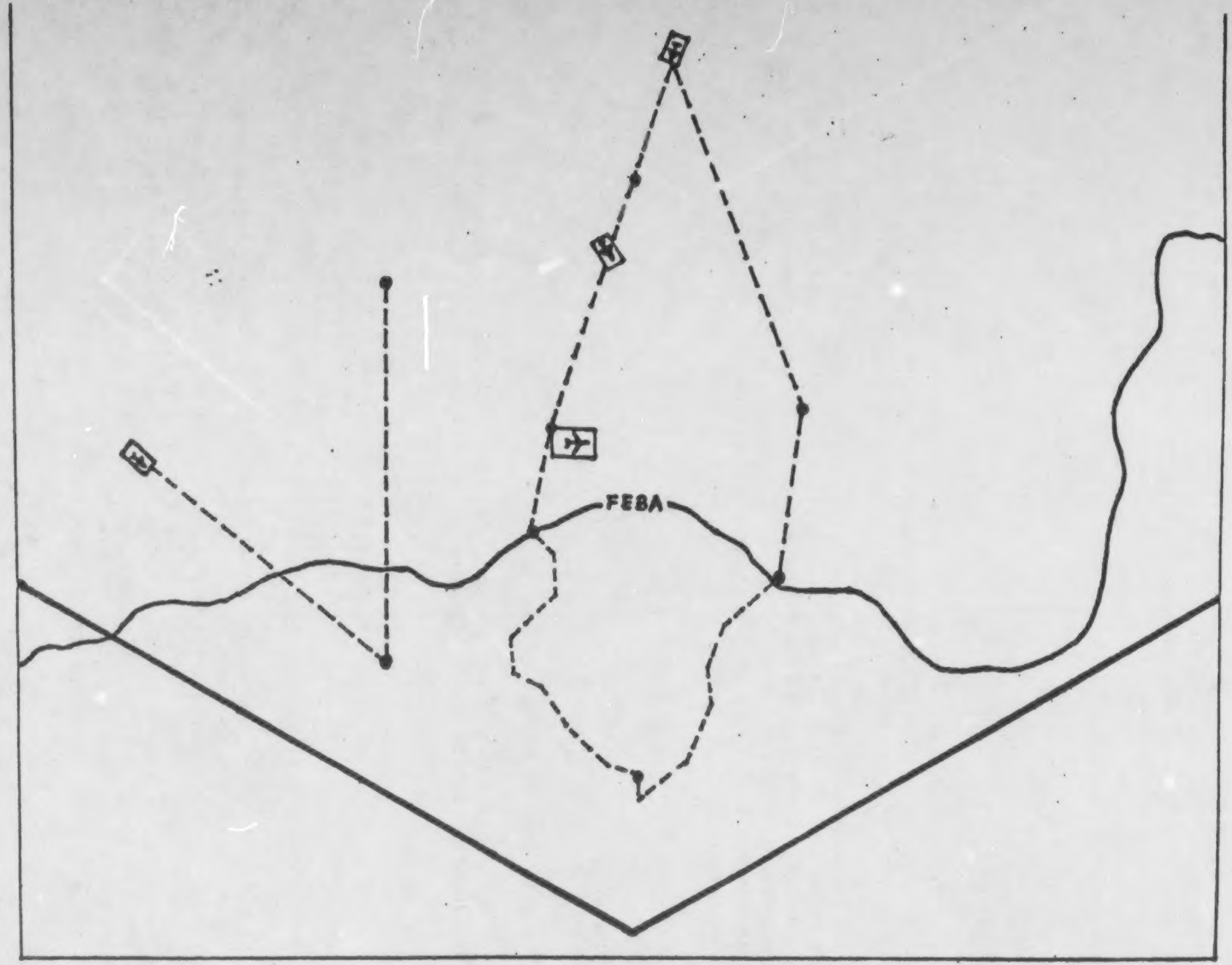


Figure 2. Scenario used in operator performance evaluation.

$$EFF = \frac{I}{Ti} \cdot \frac{1}{TP} \cdot \frac{1}{Ti}$$

where;

$ti = tit - te$

yields a measure of EFF.

The number of resources spent per threat is represented in the model as follows:

$1 + Nr - Ni$

where Nr = number of resources spent, and

Ni = number of incapacitated threats.

To incorporate this unit measure of resource effectiveness into the model the reciprocal of the difference of number of resources spent and number of incapacitated threats plus a constant of one is multiplied by the product of the incapacitated threats, threat penatrators, and time to incapacitate threats ratios. Therefore, the model assumes the form:

$$EFF = \frac{I}{Ti} \cdot \frac{1}{TP} \cdot \frac{1}{Ti} \cdot \frac{1}{RS}$$

where;

$RS = 1 + Nr - Ni$

The third variable is the effect of asset value or asset threat weight which impacts on the value of threat penatrators (TP).

A preliminary analytical evaluation will beconducted to determine the relationship between the asset threat weight and TP. In order to give the asset threat weight some value the following assumption is made: D is equal to the value of four minus the asset threat weight, where the asset threat weight assumes the values of one to three. The model now is shown as:

$$EFF = \frac{I}{Ti} \cdot \frac{1}{TP(D)} \cdot \frac{1}{Ti} \cdot \frac{1}{RS}$$

The model is designed to evaluate one threat.

To generalize across x number of threats an EFF value representing the average of the sum across x number of threats is produced. The final form of the proposed model is:

$$EFF = \sum_{i=1}^{x} \frac{\left[\frac{I}{Ti} \cdot \frac{1}{TP(D)} \cdot \frac{1}{Ti} \cdot \frac{1}{RS}\right]}{Tx} \Bigg| .$$

In summary, the model provides the researcher with a means for determining the system's performance on a given scenario. The system's effectiveness measure is used as the "yardstick" or baseline for evaluating operators' performance on the scenario. Operators' (actual) actions are then evaluated against the system's (model) actions. An evaluation of mdoel versus actual performance permits review and analysis of operators' strengths and weaknesses. Subsequently, training material can be developed to remove noted operators' weaknesses.

The proposed model is only in its developmental stages. Numerous revisions and refinements are anticipated. The evolving nature of the model will reflect enhancements that improve both the validity and reliability of the effectiveness measure.

GENERALIZABILITY AND FUTURE RESEARCH

The methodology discussed in this paper is being developed specifically for the PATRIOT Air Defense Weapon System. It is anticipated, however, that this methodology of evaluating operator performance be expanded to encompass additional systems requiring operator performance on radar systems such as those found in the Army, Air Force, Navy, Marines, and Coast Guard. Generalizing the methodology to the Federal Aviation Association's training of Air Traffic Controllers also is envisioned.

This methodology, although in its preliminary stages of development, offers promising possibilities for the evaluation of operator performance.

References

Stephens, C. E., Software Spoken Here., Air Defense Magazine, April 1978, pp 21-25

CRITICAL PERFORMANCES OF BATTALION COMMAND GROUPS*

Ira T. Kaplan and Herbert F. Barber
Army Research Institute for the Behavioral and Social Sciences
Field Unit, Fort Leavenworth, Kansas

Abstract

The behavior of 23 battalion command groups was investigated in a simulated combat environment provided by the Combined Arms Tactical Training Simulator (CATTS). Thirteen mechanized groups performed a covering force operation followed by an attack, and ten non-mechanized groups performed a defense and an attack. Their performance was rated on items derived from the subtasks of the battalion command group ARTEP (Army Training and Evaluation Program). Fifteen subtasks were identified as critical, because they or their elements were both low-rated and highly correlated with ratings of overall effectiveness.

The four missions observed in this investigation were markedly different with respect to subtask criticality. All but one of the 15 critical subtasks were identified in the covering force mission, five subtasks were critical in the mechanized attack, one in the defense, and one in the non-mechanized attack.

Rater reliability was low. The coefficient of reliability was only .22 for scores from a single rater. It increased to .55 when the scores from four or five raters were averaged. The differences among ratings of the same command group by different observers were significant beyond the .001 level. These results indicate a need for further research to develop more objective measures of command group performance.

---

*The views expressed herein are the authors' and are not necessarily endorsed by the U.S. Army.

INTRODUCTION

BACKGROUND

In recent years, time and resource constraints have provided increased impetus for the development of more efficient military training systems. The Army Training and Evaluation Program (ARTEP) and the various battle simulations are prime examples of such systems.

The Combined Arms Training Developments Activity (CATRADA) at Fort Leavenworth, Kansas is responsible for developing ARTEPs and battle simulations for battalion, brigade, and division command groups.[1] At the battalion level, CATRADA has developed the Command Group/Staff Module of ARTEP 71-2,[2] which specifies the training objectives for the battalion commander and his staff. In addition, CATRADA is developing four different battle simulations for training battalion command groups. Each simulation has its own unique capabilities and limitations. Pegasus is a manual control system for battalion and brigade CPX's (command post exercises). BATTLE (Battalion Analyzer and Tactical Trainer for Local Engagements) is played on a terrain board with the aid of a mini-computer. CAMMS (Computer Assisted Map Maneuver System) exercises battalion, or brigade and battalion, command groups via terminals linked by telephone to a large, time-shared computer. CATTS (Combined Arms Tactical Training Simulator) is the most realistic and the most completely automated battle simulation available for training battalion command groups. It is supported by a large, dedicated computer and a full-time controller staff. CATTS is permanently located at Fort Leavenworth, but a remote version is being developed that will be able to provide exercises at a unit's home station. These battle simulations and the command group ARTEP are subsystems within a larger system for training battalion commanders and their staffs. Courses taught in Army schools, and CPX's and field exercises conducted by the units themselves are also elements of the command group training system.

The systems approach to training, as described in the Instructional Systems Development Model,[3] is the approved methodology to be followed in the development of military training systems. A simple outline of the systems approach, similar to Eckstrand's model,[4] will serve to place the present research in the context of a systems approach to training development.

---

[1] Battle Simulations and the ARTEP. Combined Arms Training Developments Activity, Ft Leavenworth, KS, 18-461, November 1977.

[2] Army Training and Evaluation Program for Mechanized Infantry/Tank Task Force, No. 71-2, Headquarters, Department of the Army, Washington, D.C., 17 June 1977.

[3] Interservice Procedures for Instructional Systems Development. TRADOC Pamphlet 350-30. Fort Monroe, VA: U.S. Army Training and Doctrine Command, 1975.

The development of a training system can be described as a seven stage process:

1. Define the training objectives.
2. Develop measures of performance.
3. Derive the training content.
4. Design training methods and materials.
5. Conduct training.
6. Evaluate trainee performance.
7. Provide feedback to modify content, methods and materials.

The relationships among these stages are diagrammed in the flow chart in Figure 1, and their relevance to the development of command group training is elaborated below.

The Command Group/Staff Module of the ARTEP defines the objectives of the battalion command group training system. The module comprises 12 tasks, which are broken down into a total of 61 subtasks, a brief statement of the conditions under which each task and subtask is performed, and a general description of the performance standards for each task and subtask. The tasks include such actions as Develop a plan based on mission, Prepare and organize the battlefield, See the battlefield during the battle, and Concentrate/shift combat power.

The measurement instruments developed for this research are questionnaires answered by experienced evaluators who observe the command group's behavior in a simulated combat environment. The questionnaire items were derived from the command group ARTEP and from previous research on the performance of battalion command groups in simulated combat.[5]

---

[4] Eckstrand, G. A. Current Status of the Technology of Training. AMRL Document, Technical Report 64-86, September 1964.

[5] Barber, H. F. and Kaplan, I. T. Battalion Command Group Performance in Simulated Combat. ARI Technical Paper. In press.

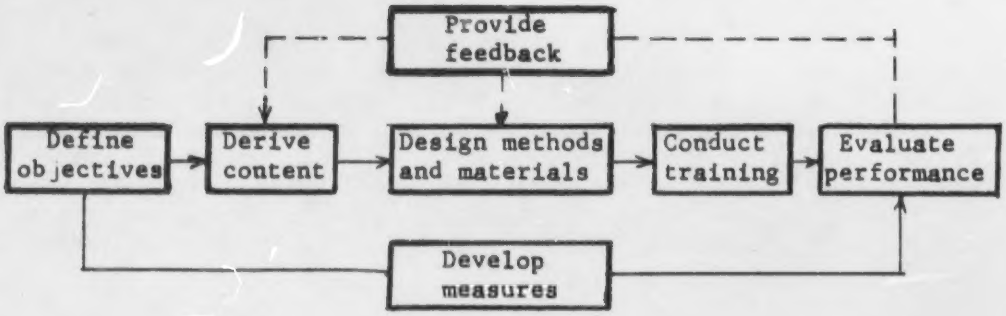

Provide feedback
Define objectives
Derive content
Design methods and materials
Conduct training
Evaluate performance
Develop measures

Figure 1. A systems approach to training development.

1327

The training content should also be derived from the objectives, because it should be directed toward satisfying those objectives. In the case of verbal material, the content is expressed in the topics, subjects and substance of books, lectures and courses. For battle simulations, the content is specified by the exercise scenarios, which determine the terrain on which the simulated battle is fought, the mission that the command group is assigned, and the ARTEP tasks and subtasks that must be performed to accomplish the mission.

The methods and materials for training battalion commanders and staffs include lectures, field manuals, exercises, and battle simulations. These simulations are used both to conduct training and, with the aid of performance measures based on the ARTEP, to evaluate the effectiveness of previous training. The feedback that an individual command group receives about its weaknesses enables it to modify its own training program to address those weaknesses. Knowledge of the weaknesses common to many command groups enables the Army to improve the total training system.

PURPOSE

The present investigation was concerned with the second and seventh steps in the system development process: deriving performance measures from the training objectives, and providing feedback to improve training content, methods and materials. This effort contributed directly to two aims of the ARI Field Unit's research on command and control training: to identify critical command group performance requirements at battalion and higher command levels, and to develop ways of measuring these performances through the use of battle simulations. More specifically, the purposes of this investigation were as follows:

1. To develop a battery of performance measures that can be used to evaluate (a) specific command-group performances, and (b) the overall effectiveness of individual staff members and the command group as a whole.

2. To identify the command group performances that are most important for training in terms of (a) low performance ratings and (b) high correlations with overall effectiveness ratings. In other words, to identify those performances that are deficient and/or strongly related to overall effectiveness.

METHOD

BATTALION COMMAND GROUPS

Data were collected from 13 mechanized and 10 non-mechanized units stationed in the Continental United States. The mechanized units included five infantry, three armor and five cavalry battalions. The non-mechanized units were all infantry battalions. Three mechanized infantry battalions were National Guard units, and all the other units were Active Army, as shown in Table 1.

The battalion command group typically comprised the commander, S1, S2, S3, S4, an air liaison officer (ALO), a fire support coordinator, an operations sergeant, intelligence sergeant, assistant S2 and/or S3 air, fire support NCO, and one or two radio/telephone operators.

EXERCISES

Each command group was observed during the performance of two missions in the CATTS facility at Fort Leavenworth. The particular missions assigned to a group depended on the type of unit it commanded, as shown in Table 2. Mechanized units performed a covering force operation followed by a daylight attack as part of a larger force. Non-mechanized units first performed a defense and then a non-illuminated, non-supported night attack. Differences in mobility and probable real-world missions determined the types of missions assigned. Active Army groups conducted their two missions during a three-day exercise. National Guard groups performed one mission per day during a two-day weekend exercise.

SIMULATION SYSTEM

The battlefield environment was simulated by the Combined Arms Tactical Training Simulator (CATTS), which provides a computer-driven exercise to train maneuver-battalion commanders and their staffs in the control and coordination of combined-arms operations. It simulates the actions of units in combat, moves elements on and above the battlefield, calculates intervisibility and detection between forces, weapon-to-target ranges and the effects of weapons employment, and it maintains the status of personnel, equipment, ammunition and fuel for friendly and enemy forces. Speed of movement, line of sight, and weapons effects are affected by changes in weather, terrain contour and soil type, suppressive fires, and personnel and equipment status. The CATTS exercise is

TABLE 1
Number of Command Groups

Type of Unit	Active Army	National Guard
Mechanized		
Infantry	2	3
Armor	3	0
Cavalry	5	0
Non-mechanized		
Infantry	10	0

TABLE 2
Type of Mission

Type of Unit	Mission 1	Mission 2
Mechanized	Covering Force	Attack
Non-mechanized	Defense	Attack

conducted in a real-time, free-play mode. Within the prescribed tactical situation, the battalion commander can employ his assets in any manner he deems appropriate. The only constraints are the assets available to the battalion and the actions of the enemy commander.

In this research, the command group occupied a simulated tactical operations center (TOC), except for the S1 and S4 who were in another area designated as the combat trains. The players (the battalion command group) in both areas were provided with communications equipment normally found in a maneuver battalion. They could communicate with higher, lower and adjacent units (played by controllers) in any manner consistent with Army procedure and with the simulated location of the various units: face to face, by telephone or radio, and by written message. Figure 2 illustrates the communications among the players, the controllers and the computer. Most communication took place over radio and telephone. The battalion command group had seven radio nets (actually hard-wired) with appropriate alternate frequencies. The nets included the brigade command net, the brigade intelligence net, the brigade administration/logistics net, the battalion command net, the fire support net, and the air support net. In addition to the radio nets, the command group also had a RATT (radio-teletype) unit and field telephones, when appropriate. The sounds of enemy jamming, battle, and engine and generator noise were generated during the exercise to add to the realism of the experience.

CONTROLLERS

A team of controllers, permanently assigned to CATTS, mediated between the players and the computer. The control group included a chief controller, who played the role of brigade commander, a brigade S1/S4 controller, who also played the roles of service-support-unit commanders and executive officers, a brigade S2/S3 controller, four maneuver- and supporting-unit commanders, a fire support controller, two forward observers, a direct air support controller (DASC), and a threat controller. The DASC was played by a different Air Force officer each time. The monitor was an adjunct member of the control group, who observed the command group during the exercise and provided feedback to the players during the post-game critique. This position was a rotating assignment among faculty members of the Command and General Staff College who had served as battalion commanders or staff members and held the rank of lieutenant colonel.

The members of the control group are listed in Table 3. Three controllers, identified as interactors, input orders to the computer through a control console: (1) the command and control interactor relayed orders

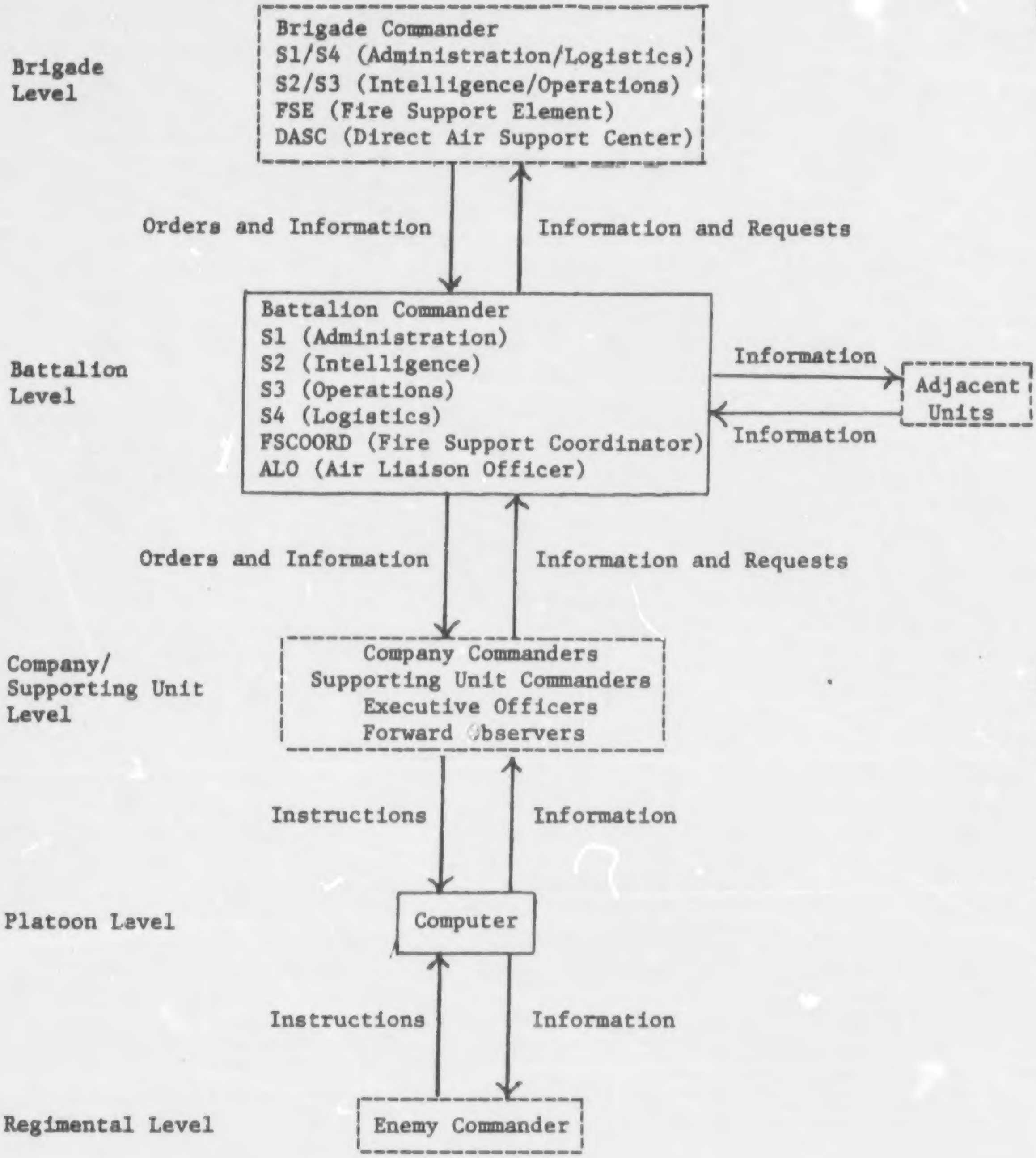

Brigade Level
Brigade Commander
S1/S4 (Administration/Logistics)
S2/S3 (Intelligence/Operations)
FSE (Fire Support Element)
DASC (Direct Air Support Center)
Orders and Information
Information and Requests
Battalion Level
Battalion Commander
S1 (Administration)
S2 (Intelligence)
S3 (Operations)
S4 (Logistics)
FSCOORD (Fire Support Coordinator)
ALO (Air Liaison Officer)
Information
Adjacent Units
Information
Orders and Information
Information and Requests
Company/ Supporting Unit Level
Company Commanders
Supporting Unit Commanders
Executive Officers
Forward Observers
Instructions
Information
Platoon Level
Computer
Instructions
Information
Regimental Level
Enemy Commander

Figure 2. Communication between controller and player positions in CATTS. Controller positions are inclosed by broken lines.

from the battalion command group to the maneuver units modeled in the computer, (2) the fire support interactor input orders to the artillery and air support units, and (3) the threat interactor, working independently, controlled the enemy force. The results of simulated movements and engagements were displayed on television screens to the controllers, who transmitted relevant information to the players via radio or telephone. Except for the threat interactor and the monitor, all controllers acted the roles of higher, lower or adjacent unit personnel. In addition to their other functions, eight controllers also filled out observation forms on which they evaluated the command group's performance in the areas designated in Table 3.

PERFORMANCE MEASURES

The documentation of the observation forms that were used to evaluate the performance of battalion command groups in simulated combat is one of the principal objectives of this report. These forms constitute a source of test items that can be drawn upon by exercise directors or battalion and brigade commanders. The items can also be used by researchers to study command group training and performance. While still subject to refinement, especially with respect to increased objectivity, these observation forms were superior to those previously employed. The major improvements, which were based on extensive experience using the earlier forms, were (1) the introduction of a five-point scale for rating the performance of ARTEP subtasks and (2) the addition of many specific questions that were answered yes or no. The five-point rating scale permitted finer discriminations than the three-point scale previously used, and the yes/no questions provided more detailed information about the components or elements of subtask performance.

Four different observation forms were filled out by the evaluators:

1. A form concerned with administration and logistics was completed by the brigade S1/S4 controller.

2. Intelligence and operations forms were completed by the brigade S2/S3 controller and by the controllers who played company commanders.

3. A fire support form was completed by the fire support controller.

4. An observation form covering all the preceding areas, in somewhat less detail, was completed by the monitor.

Each observation form had two or three versions, appropriate to the mission that was played. The S1/S4, FS and monitor's forms had one

TABLE 3
CATTS Control Group

Position	Rank	Performance Observed
Chief Controller	LTC	-
Brigade S1/S4	MAJ	Administration and Logistics
Brigade S2/S3	MAJ	Intelligence and Operations
Command and Control Interactor	CPT	Intelligence and Operations
Company Controller	MAJ	Intelligence and Operations
Company Controller	CPT	Intelligence and Operations
Company Controller	CPT	Intelligence and Operations
Unit First Sergeants	SSG	-
Fire Support Interactor	CPT	-
Artillery Controller	LTC	Fire Support
Artillery Controller	CPT	-
Artillery Controller	SSG	-
Division Air Support Center	CPT to LTC	-
Threat Interactor	CPT	-
Monitor	LTC	ARTEP Subtasks

version for the covering force or defense and another version for the mechanized or non-mechanized attack. The intelligence and operations form had one version for the covering force, another for the defense, and a third for the mechanized or non-mechanized attack. Most of the items on a given form were the same in both or all three versions, but some items were unique to the mission.

The performances evaluated by the brigade S1/S4 controller included subtasks and elements of subtasks concerned with providing supplies and maintaining equipment before and during the battle, (Subtasks 3J, 3K, 9A and 9B), supporting the troops (9C), and integrating combat service support (CSS) into the scheme of maneuver (9D). The S1/S4 controller also rated the overall effectiveness of the battalion S1 and S4 in comparison with those of previous command groups.

The brigade S2/S3 controller and the company commanders evaluated subtasks and elements concerned wtih intelligence preparation of the battlefield (1B, 2A, 2B, 2C, 2D), analyzing friendly capabilities (1D), selecting routes and positions (1F, 1G, 1H), organizing for combat (3C), communicating plans and orders (3D, 3F, 3G), seeing the battlefield during the battle (5B, 5C, 5D), troop leading during the battle (11A), communicating changes (6B), concentrating combat power (9A, 8B, 8C, 8D), and reacting to enemy electronic warfare (10A, 12A). They also rated the overall effectiveness of the battalion S2, S3 and the command group as a whole in comparison with previous S2's, S3's and command groups.

The fire support controller evaluated the fire support plan (1I), priority of fires (1J), fire support coordination (1L), modification of the fire support plan during the battle (7A), and the overall effectiveness of the battalion fire support element in comparison with that of previous groups.

The monitor evaluated all the above types of performance by rating subtasks of the command group ARTEP. The only subtasks he did not rate were those that were not played in the exercise and those he could not observe. He also evaluated the degree to which the mission was accomplished, and the overall effectiveness of the battalion commander and the command group as a whole. Since the monitor did not have as much experience observing command groups as the permanent controllers, he did not rate them in comparison with previous groups, but used the five-point performance scale instead.

The scales that were used to rate performance, overall effectiveness and mission accomplishment are listed below:

The alternative responses on the five-point performance scale were defined as follows:

1 - Completely overlooked, forgotten
2 - Major deficiencies
3 - Minor deficiencies
4 - Satisfactory
5 - Excellent

The overall effectiveness of the command group as a whole and of its individual members were rated in comparison with previous command groups and their members on the following scale:

1 - One of the worst
2 - Worse than average
3 - Average
4 - Better than average
5 - One of the best

The alternative responses for the monitor's evaluation of mission accomplishment were:

1 - Failed to accomplish any part of the mission
2 - Failed to accomplish most of the mission
3 - Accomplished about half of the mission
4 - Accomplished most of the mission
5 - Accomplished all of the mission

DATA ANALYSIS

The primary objectives of the data analysis were to identify those performances that were deficient and those that were highly correlated with ratings of overall effectiveness. Performances were designated as deficient when their average scores were less than or equal to the midpoint of the rating scale. For performances rated on a five-point scale, the criterion score was 3.0. For items answered yes or no, the criterion score was 50% yes. Most of the yes/no items were worded so that a yes response signified a correct performance in the rater's judgement, but a few questions were phrased so the proper behavior was indicated by a no response. Thus, the general definition of deficient performance for yes/no items was an average score less than or equal to 50% correct.

A correlation was considered high when it was statistically significant at the .01 level by the one-tailed test. Pearson r's were computed for the scaled items and point biserial correlations for the yes/no items. The .01 criterion of significance was chosen in preference to the less stringent .05 level, because of the large number of correlation coefficients (over 2,000) that were computed. About 100 would be significant by chance at the .05 level, versus about 20 at the .01 level. The one-tailed test was used because only a positive relationship was expected between correct performance and overall effectiveness.

One problem with using correlation as a measure of criticality is that the size of the correlation between two variables decreases when the range of either variable is restricted.[6] For example, if every S4 determined the status of equipment before the battle, there would be no correlation between performance of this task and the S4's overall effectiveness ratings, even though failure to perform the task might have harmful consequences. In fact, however, lack of variation was not a serious problem in this study, because the typical case of restricted range occurred when the task was usually performed correctly. It can be argued that a task which is performed correctly by a given population is not important for training in that population. Scandura[7] made a related point, when he wrote that professional competence does not have to be analyzed to the level of elementary skills. For example, all accountants can add, so arithmetic ability is not an important variable for distinguishing among individuals within the population of trained accountants. Similarly, the tasks that were usually performed correctly in the present investigation are not critical for training incumbent command groups. This argument is consistent with the ARTEP philosophy, which advocates training to correct deficiencies.

The data for each of the four missions were analyzed separately. First, mean scores were calculated for the items that were rated by several observers, i.e., the intelligence and operations items that were rated by the brigade S2/S3 and the company commanders, the overall effectiveness ratings for the battalion S2 and S3 rated by the same controllers, and the overall effectiveness ratings for the whole command group provided by the same controllers plus the monitor. These ratings were averaged over observers to obtain a mean score for the command group on each item. Then the scores for every command group on all four observation forms (administration and logistics, intelligence and operations, fire support, and the monitor's form) were analyzed to produce the two desired measures of performance for each item:

---

[6] Welkowitz, J., Ewen, R.B., and Cohen, J. Introductory Statistics for the Behavioral Sciences, 2d Ed. New York: Academic Press, 1976.

[7] Scandura, J.M. Structural approach to instructional problems. American Psychologist, 1977, 32, 33-54.

1. The mean rating averaged over all the command groups that played a given mission.

2. The correlations between performance on the item and the overall effectiveness ratings, calculated across all the command groups that played a given mission.

In addition to the analysis of ratings and correlations for every item on the observation forms, analyses of variance were performed to compare the relative difficulty of the four different missions, and measures of inter-rater reliability were computed for the items that were rated by several observers.

RESULTS

The results of this investigation are described below under three major headings: (1) performance deficiencies, (2) performances correlated with overall effectiveness, and (3) performances that were both deficient and correlated with overall effectiveness. Under each heading, four sets of command group performances are considered in turn: (1) administration and logistics, (2) intelligence and operations, (3) fire support and (4) the subtasks rated by the monitor. Within each of these sets, the results are presented for each of the four missions: (1) covering force, (2) mechanized attack, (3) defense and (4) non-mechanized attack.

PERFORMANCE DEFICIENCIES

A given performance was considered deficient when it was rated incorrect for 50% or more of the command groups - in the case of items answered yes or no, or when its average score was less than or equal to 3.0 - in the case of items rated on a five-point scale. All the deficient performances are listed in Tables 4 through 7.

A performance that was deficient in one mission was not necessarily deficient in another. To make the deficient performances stand out more clearly, satisfactory performance scores are not included in Tables 4 through 7. The mean score for each item is entered in the column that corresponds to the mission in which the performance was deficient. The entries with decimal points are mean scores on the five-point scale.

Percentage entries indicate the percent of yes responses to items that were answered yes or no. The entry N/A (not applicable) means that there is no score because the item did not appear on the observation form for that mission.

Items that correspond to ARTEP subtasks in Tables 4 through 6 are identified by the subtask label in parenthesis after the item. The elements of a subtask are listed before it. When elements of a subtask were deficient, but the overall subtask was not, the subtask label is given before the elements.

Administration and Logistics. Table 4 shows that the deficiencies in administration and logistics had to do with providing supplies and equipment during the battle (Subtask 9A), and integrating combat service support into the scheme of maneuver (Subtask 9D). Four performances were deficient in the covering force operation (Items 6d, 6e, 9b and 9c on the observation form), none in the mechanized attack, one in the defense (6d), and one in the non-mechanized attack (6g). Obviously, the specific deficiency in the use of transportation assets (9b) contributed to the more general weakness in the integration of CSS into the scheme of maneuver (9c) in the covering force.

Intelligence and Operations. All the items that were deficient according to the average ratings of the S2/S3 and company commander controllers are listed in Table 5. There were deficiencies in eight of the ten categories of items; the two exceptions were: B. Friendly considerations and, C. Organize for combat. Because the items within each category varied from one mission to another, a given item on the observation form for one mission generally had a different number on another form. For ease of reference, therefore, the items in Table 5 have been renumbered consecutively within each category.

Six items were deficient in at least three of the four missions:

A7. The intelligence collection plan was not properly prepared.

G1. The command group sometimes made unwarranted assumption that all team commanders were monitoring their radios for changes. (This was an instance where a yes response indicated the incorrect behavior.)

I1. There was too much radio communication. (Another case where yes meant wrong.)

I2. There were security violations during radio traffic. (A third such case.)

TABLE 4
Deficiencies in Administration and Logistics

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
6. In providing supplies and equipment to arm and fuel the system during the battle: (9A)				
d. Did the S4 coordinate with the S2 so he knew the enemy's capabilities?	50%	-	33%	-
e. Did the S4 keep his higher appropriately informed of his activities?	45%	-	-	-
g. Did the S4 effectively utilize his direct support assets?	-	-	-	44%
9. In terms of integrating CSS into the scheme of maneuver:				
b. Were transportation assets used to fit movement of CSS resources to the scheme of maneuver?	46%	-	-	-
c. How effectively was CSS integrated into the scheme of maneuver? (9D)	3.00	-	-	-

TABLE 5
Deficiencies in Intelligence and Operations

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
A. Intelligence preparation of the battlefield				
1. Was the enemy's scheme of maneuver and fire support identified?	41%	48%	-	-
2. Was the enemy's ability to attack by air identified?	45%	N/A	46%	N/A
3. Was the enemy's nuclear capability identified?	32%	N/A	36%	N/A
4. Was the enemy's chemical capability identified?	39%	N/A	44%	N/A
5. Overall, how well did the command group identify critical combat information and intelligence? (1B, 2A)	2.76	2.97	-	-
6. Were all GSR elements effectively utilized?	-	10%	-	-
7. Was the TF intelligence collection plan properly prepared, and did it reflect analysis by the battalion S2 of tasking responsibilities?	2.67	2.47	2.74	2.53
8. Overall, how well did the command group determine combat information and intelligence shortfalls and aggressively gather information from all available/appropriate sources? (2B)	2.71	2.73	-	-
9. Was relevant information from higher headquarters and adjacent units disseminated to company commanders (e.g., minefields)?	2.85	2.98	-	-

N/A = Not applicable.

TABLE 5 (Continued)
Deficiencies in Intelligence and Operations

	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
10. Were company commanders given an estimate of specifically what they would be facing?	36%	-	-	-
11. Overall, did the command group disseminate combat information and intelligence that was event-oriented and usable to the recipient? (2D)	2.87	-	-	-
D. Communicate/coordinate plans and orders.				
1. Were company commanders given instructions on actions to be performed if jamming occurs?	-	-	-	49%
2. Were effective alternate means of communication developed in case of lost commo?	32%	44%	-	-
3. Was wire utilized as an effective means of communication?	N/A	35%	N/A	-
4. Did the command group develop a communication plan which satisfies the communications requirements of the specific mission, provides for COMSEC, specifies alternative means of communication, and insures operation of MIJI plan? (3F)	2.79	2.74	-	-
5. Did all elements understand what they were to do without extensive questioning?	46%	-	-	-
6. Did the operation order contain enough information for attached units?	2.86	-	-	-

N/A = Not applicable.

TABLE 5 (Continued)
Deficiencies in Intelligence and Operations

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
7. Was sufficient time allowed to task force elements for their troop leading procedures?	44%	44%	-	-
8. Overall, were the orders appropriate, clear, concise, and did they contain essential information; were they issued so as to allow TF elements maximum time to go through troop leading procedures; and were they coordinated with proper agencies? (3G)	2.79	3.00	-	-
E. See the battlefield during the battle.				
1. How well did the command group disseminate information and intelligence that was event-oriented, usable to the recipient, accurate, and within a time frame which permitted the recipient to react? (5D)	3.00	-	-	-
F. Troop lead during the battle. (11A)				
1. Were all attached combat units adequately controlled/monitored during the conduct of the exercise?	2.94	-	-	-
G. Coordinate/communicate changes. (6B)				
1. Did the command group sometimes assume all commanders were monitoring radios for changes?	56%	64%	56%	66%
H. Concentrate/shift combat power.				
1. How well did the command group read the battlefield & determine the precise place & time for maximum combat power needed to be employed? (8A)	2.60	-	-	-

TABLE 5 (Continued)
Deficiencies in Intelligence and Operations

	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
2. When the enemy committed itself, did the command group adequately redeploy forces?	2.43	N/A	-	N/A
3. Were tactical decisions made consistent with the time-distance relationship?	39%	-	-	-
4. Overall, how well did the command group concentrate its organic/attached/DS assets according to the weapons capabilities and movement of the enemy force? (8B/C)	2.76	-	-	-
5. Overall, how well did the command group direct organic/supporting forces to conduct economy of force operations in the thinly held areas (when concentrating combat power)? (8D)	2.80	3.00	-	-
I. Enemy EW Considerations.				
1. Was there too much communication?	88%	73%	84%	-
2. Did security violations occur during radio traffic?	54%	61%	50%	61%
3. Overall, how well did the command group adhere to communications and electronic security measures? (10A)	2.81	2.81	-	-
4. Was a MIJI report promptly submitted to higher headquarters using secure means of communication?	-	40%	-	-

N/A = Not applicable.

TABLE 5 (Continued)
Deficiencies in Intelligence and Operations

	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
5. Did the command group direct a switch to spare frequency as a last resort using proper authentication techniques?	23%	36%	34%	43%
6. Overall, how well did the command group recognize and react to enemy electronic warfare? (12A)	2.82	2.85	-	-
J. Other.				
1. Was the sufficient intra-staff coordination between 2/3 and 1/4?	41%	38%	-	-
2. Was there sufficient coordination between NCS and 2/3?	43%	49%	-	-
3. How well did the command group apply the time-distance relationship while maneuvering Task Force elements?	2.86	-	-	-
4. Did the Task Force maneuver elements become decisively engaged because of battalion action?	56%	N/A	N/A	N/A
5. What was the size of the battalion reserve?	23%	-	N/A	N/A

N/A = Not applicable.

NCS = Net Control Station.

I5. Spare frequencies were not used correctly.

J5. The battalion reserve was too large. The percentage for this item indicates the size of the battalion reserve, which should have been about 10%, in the judgement of the raters.

The largest number of deficiencies occurred in the covering force operation (33) and the next largest in the mechanized attack (22). There were relatively few deficient performances in the defense (9) or in the non-mechanized attack (6). The greatest concentrations of deficiencies were in categories:

A. Intelligence preparation of the battlefield.

D. Communicate/coordinate plans and orders.

H. Concentrate/shift combat power.

I. Enemy EW considerations.

Categories A, D and I were weak in the covering force and mechanized attack; Category H, mainly in the covering force.

Fire Support. The deficiencies in fire support are labelled in Table 6 as they were on the observation form. Only Item 1a, planning the utilization of heavy mortars, was deficient in several types of mission. All the other deficiencies, which were related to determining the initial priority of fires (Subtask 1J) and modifying the fire support plan (Subtask 7A) were confined to the covering force. Items 4a and 4b were parts of the overall performance encompassed by Item 4c.

Monitor's Ratings. The subtasks evaluated by the monitor are labelled in Table 7 as they were on his observation form and in the command group ARTEP. Only four subtasks were deficient in three of the four missions: 2C (Analyze enemy), 2D (Disseminate critical intelligence), 3G (Communicate/coordinate plans and orders), and 12A (React to enemy electronic warfare). However, twelve subtasks, including 2C, 2D and 3G, were deficient in both the covering force and the mechanized attack; these were primarily the subtasks concerned with intelligence before (2A, 2B, 2C and 2D) and during (5A, 5B, 5C and 5D) the battle, and with managing combat service support (9A, 9B and 9C). Altogether 19 items were deficient in the covering force and 15 in the mechanized attack, compared to just three in the defense and three in the non-mechanized attack.

TABLE 6
Deficiencies in Fire Support

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
1. Plan use of organic/attached and non-organic fires. (1I)				
a. Did fire plan effectively utilize heavy mortars?	20%	-	40%	50%
2. Determine priority of fires.(1J)				
a. Was priority of fires given to appropriate TF element(s) to support scheme of maneuver?	50%	-	-	-
b. Was suppression of fires considered?	50%	-	-	-
4. Modify fire support plan.				
a. During the battle was priority of fires supporting new scheme of maneuver immediately communicated to supporting and supported units?	50%	-	-	-
b. Were requests for immediate fire support received and assigned to appropriate fire support agencies?	50%	-	-	-
c. Overall, how well did the command group perform relative to the standard? (7A)	2.90	-	-	-

TABLE 7

Deficiencies Observed by the Monitor

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
1. Develop plan based on mission.				
B. Identify critical intelligence.	-	2.73	-	-
I. Plan use of organic/attached and non-organic fires.	2.75	-	-	-
2. Initiate intelligence preparation of the battlefield.				
A. Identify critical intelligence.	2.55	2.67	-	-
B. Gather critical intelligence.	2.55	2.83	-	-
C. Analyze enemy.	2.64	2.92	-	3.00
D. Disseminate critical intelligence.	2.73	2.91	2.88	-
3. Prepare and organize the battlefield.				
E. Plan organic, attached and non-organic supporting fires and determine priority.	-	3.00	-	-
F. Develop a communication plan.	-	2.71	-	-
G. Communicate/coordinate plans and orders.	2.73	2.73	2.83	-
I. Employ active/passive security measures.	2.60	-	-	-

TABLE 7 (Continued)
Deficiencies Observed by the Monitor

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
5. See the battlefield during the battle.				
A. Identify critical intelligence.	2.75	2.70	-	-
B. Gather critical intelligence.	2.92	2.80	-	-
C. Analyze enemy.	2.42	3.00	-	-
D. Disseminate critical intelligence.	3.00	2.89	-	-
7. Employ fires and other combat assets.				
A. Modify fire support plan.	2.92	-	-	-
B. Employ fires.	3.00	-	-	-
8. Concentrate/shift combat power.				
A. Determine critical place and time.	2.92	-	-	2.83
B/C. Concentrate/shift combat power.	2.83	-	-	-
9. Manage combat service support assets.				
A. Arm and fuel the systems.	3.00	3.00	-	-
B. Fix the system.	3.00	2.89	-	-
C. Support the troops	3.00	3.00	-	-
D. Integrate CSS into scheme	-	-	-	-

TABLE 7 (Continued)
Deficiencies Observed by the Monitor

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
12. React to situations requiring special actions.				
A. React to enemy electronic warfare.	3.00	-	3.00	2.00

1350

Summary. The percentage of items that were judged deficient in each type of mission is shown in Figure 3. The four percentages for each mission refer to (1) administration and logistics rated by the S1/S4 controller, (2) intelligence and operations rated by the S2/S3 and company commander controllers, (3) fire support rated by the fire support controller, and (4) subtasks rated by the monitor. Every evaluator reported the greatest percentage of deficiencies in the covering force operation. The next highest percentages of deficiencies were in the mechanized attack, specifically in the intelligence and operations items and in the monitor's ratings. There were relatively few deficiencies in the defense and the non-mechanized attack.

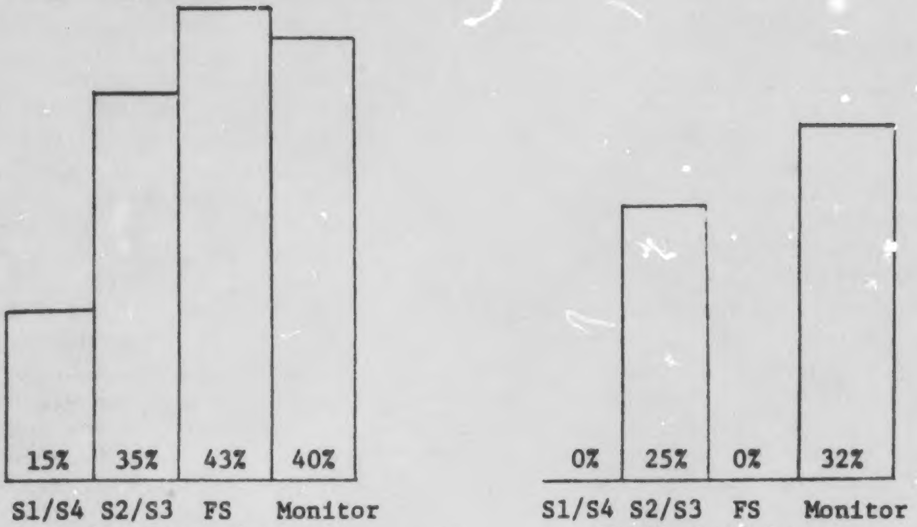
The preceding results indicate that the mechanized attack was performed better than the covering force and that both non-mechanized missions were performed better than the mechanized missions. These relationships were generally supported by further analysis of the data. Table 8 summarizes the mean scores on the ARTEP subtasks evaluated by each rater for each mission. A higher score means the subtask was performed better. An analysis of variance was done for each set of ratings. In Kirk's[8] terminology, it was a split plot factorial design with repeated measures over two factors, mission (first versus second) and subtasks, with type of unit (mechanized versus non-mechanized) a grouping factor. The ANOVAs showed that the second mission was significantly better than the first ($p < .05$) for the administration and logistics ratings, and for fire support. The non-mechanized groups scored significantly higher than the mechanized ones only on the intelligence and operations subtasks ($p < .001$). Planned comparisons (t tests) showed that the mechanized attack was performed better than the covering force for all four sets of subtasks: administration and logistics ($p < .01$), intelligence and operations ($p < .05$), fire support ($p < .001$), and the monitor's ratings ($p < .05$). The non-mechanized attack was better than the defense only for fire support ($p < .001$). Because the attack was always the second mission, it is not possible to decide whether it was an easier mission to perform or whether the command groups improved with practice from the first mission to the second. It is probable, however, that the covering force was more difficult than the defense, because those two missions were always performed first.

CORRELATES OF EFFECTIVENESS

Eight measures of effectiveness were obtained in this investigation: the overall performance of the S1 and of the S4 were rated by the S1/S4

---

[8] Kirk, R. E. Experimental Design: Procedures for the behavioral sciences. Belmont, California: Brooks Cole, 1968.



Covering Force Operation

Mechanized Attack

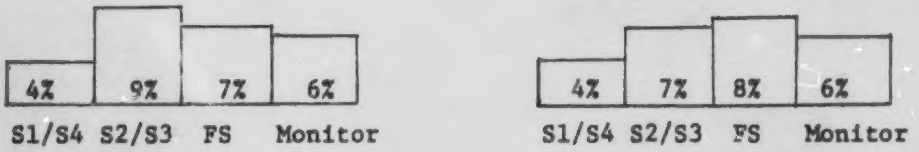


Figure 3. Percentage of items rated deficient by each rater for each type of mission.

TABLE 8
Mean Ratings on ARTEP Subtasks

Item	Mechanized		Non-Mechanized	
	CFO	Attack	Defense	Attack
Administration and logistics	2.9	3.7	3.4	3.7
Intelligence and operations	2.9	3.1	3.6	3.5
Fire support	3.2	3.9	3.4	4.1
Monitor	3.1	3.4	3.2	3.3

controller, the S2 and S3 were rated by the S2/S3 controller and the company commander controllers, the utilization of fire support assets (FS) was rated by the fire support controller, the battalion commander (Cdr) and mission accomplishment (Msn) were rated by the monitor, and the overall effectiveness of the command group as a whole (CG) was rated by the S2/S3 and company commander controllers and by the monitor.

Intercorrelations Between Effectiveness Ratings. The intercorrelations among the measures of effectiveness for each mission are shown in Tables 9 through 12. Most of the correlations in Tables 9 and 10 are based on 13 cases; most of those in Tables 10 and 11 are based on 10 cases. Correlations that are significant beyond the .01 level are marked with asterisks. The statistical significance of a correlation depends on the number of cases on which it is based, as well as on its size. Therefore, a given correlation may not be significant, even though it is larger than a significant correlation that is based on more cases.

As would be expected, the rating of the command group as a whole was the measure most highly correlated with the other measures of effectiveness. Eight of the eleven significant correlations in Tables 9 through 12 involved the command group rating. The most consistently related pair of variables were the S3 and command group ratings, which were significantly correlated in all four missions. On the other hand, only one of the monitor's effectiveness ratings was significantly correlated with another effectiveness rating, viz., mission accomplishment with command group effectiveness in the defense. Probably the reason that no other monitor's ratings were significantly correlated was that every command group was rated by a different monitor. Variation in personal evaluative criteria from one monitor to another probably reduced the correlations between effectiveness ratings. The FS rating was not significantly correlated with any other measure of effectiveness, which probably reflects a lack of integration between fire support and the other command group functions.

Administration and Logistics. Table 13 shows that most of the admin/log items significantly correlated with overall effectiveness ratings were correlated with the rating of the S4. This is a plausible result, because most of the items refer to S4 functions. There were fewer significant correlations in the non-mechanized missions than in the mechanized ones, because the performance ratings were consistently high in the non-mechanized missions. As noted in the Method Section, when the range of the variables is restricted, the correlation between them is reduced.

TABLE 9
Intercorrelations Between Effectiveness Ratings for the Covering Force Operation

	S2	S3	S4	FS	Cdr	Msn	CG
S1	.24	.47	.87**	.37	.41	.14	.73*
S2		.53	.28	.27	-.61	-.11	.38
S3			.36	.37	-.21	-.38	.68*
S4				.37	.40	.27	.66*
FS					-.15	.24	.20
Cdr						.23	-.11
Msn							-.31

*$p < .01$
**$p < .001$

TABLE 10
Intercorrelations Between Effectiveness Ratings for the Mechanized Attack

	S2	S3	S4	FS	Cdr	Msn	CG
S1	.59	.59	.63	.26	.65	.21	.50
S2		.87**	.68*	.24	.39	.22	.77**
S3			.58	.32	.42	.57	.90**
S4				.39	.33	.10	.58
FS					.02	.35	.40
Cdr						.14	.46
Msn							.66

*$p < .01$
**$p < .001$

TABLE 11

Intercorrelations Between Effectiveness Ratings for the Defense

	S2	S3	S4	FS	Cdr	Msn	CG
S1	-.51	-.67	-.08	.64	.40	-.27	-.46
S2		.64	-.24	-.27	-.42	.27	.19
S3			-.06	-.18	.10	.47	.71*
S4				.53	-.04	-.15	.07
FS					.65	.12	.07
Cdr						.63	.76
Msn							.79*

*p < .01

TABLE 12

Intercorrelations Between Effectiveness Ratings for the Non-Mechanized Attack

	S2	S3	S4	FS	Cdr	Msn	CG
S1	-.36	-.14	.09	.50	-.18	-.17	.00
S2		.38	-.18	-.13	.22	.63	.17
S3			-.08	.08	.32	.14	.73*
S4				.08	.54	.65	.15
FS					.51	-.04	.18
Cdr						-.15	.42
Msn							.52

*p < .01

Inspection of the data showed that most of the admin/log subtasks or their elements were significantly correlated with overall effectiveness ratings in all four missions. The subtasks rated by the S1/S4 controller were 3J (Provide supplies) and 3K (Maintain equipment), performed in preparation for the battle; and 9A (Arm and fuel the systems), 9B (Fix the systems), 9C (Support the troops), and 9D (Integrate CSS into scheme of maneuver), performed during the battle. There were only a few cases in which one of these subtasks or its elemetns was not correlated with a measure of effectiveness. Subtask 3K was not correlated with effectiveness in the first mission (covering force or defense), when there was no maintenance to perform. Subtasks 9C and 9D were not correlated with effectiveness in the mechanized attack, when they were performed consistently well, with little variation. With the preceding exceptions, some elements of every admin/log subtask were significantly correlated with S1, S4, or command group effectiveness in all four missions.

Intelligence and Operations. About 20% of the items in this catecory were significantly correlated with the effectiveness rating for the command group as a whole (Table 14). Somewhat fewer items were correlated with the S2 and S3 ratings. There were more significant correlations in the two mechanized missions, because there was a wider range of variation in performance, as noted above.

Most of the intel/ops items were related to ARTEP subtasks and elements thereof. The following subtasks or their elements were consistently correlated with effectiveness ratings in three or in all of the four different missions:

1B, 2A. Identify critical combat information and intelligence.

2B. Gather critical combat information and intelligence.

2C. Analyze opposing force.

2D. Disseminate critical combat information and intelligence.

3C. Organize for combat.

3G. Communicate/coordinate plans and orders.

(The preceding subtasks were performed in preparation for the battle; the following were performed during the battle.)

TABLE 13
Number of Administration and Logistics Items Significantly Correlated ($p < .01$) With Effectiveness Ratings

Effectiveness Rating	Mechanized CFO	Mechanized Attack	Non-Mechanized Defense	Non-Mechanized Attack
S1	14	2	0	0
S4	16	9	12	4
CG	5	3	0	0

NOTE: The total number of items was 26 in the CFO and defense, 28 in the attack.

TABLE 14
Number of Intelligence and Operations Items Significantly Correlated ($p < .01$) With Effectiveness Ratings

Effectiveness Rating	Mechanized CFO	Mechanized Attack	Non-Mechanized Defense	Non-Mechanized Attack
S2	9	9	8	2
S3	3	18	8	8
CG	21	17	11	19

NOTE: The total number of items was 93 in the covering force, 89 in the mechanized attack, 95 in the defense and 88 in the non-mechanized attack.

5B. Gather critical combat information and intelligence.

6B. Coordinate/communicate changes.

8A. Determine critical place and time.

8B/C. Concentrate/shift combat power. (8B, in the attack/8C, in the defense or retrograde).

10A. Defeat or suppress the enemy's electromagnetic intelligence effort.

11A. Troop lead during battle.

Another item, not explicitly part of any subtask, that was significantly correlated with effectiveness ratings in all four missions, was the question: How well did the command group apply the time-distance relationship while maneuvering Task Force elements?"

Subtasks 1B and 2A, which are separate items in the ARTEP, could not be evaluated separately by the controllers.

Fire Support. One third of the items rated by the fire support controller were significantly correlated with his rating of overall fire support effectiveness (Table 15). However, none of the items were correlated with the rating of command group effectiveness, which is consistent with the low correlation between FS and CG ratings mentioned earlier. Examination of the results showed that all four subtasks rated by the fire support controller were significantly correlated with his rating of fire support effectiveness in three of the four missions. In the non-mechanized attack, when the subtasks were performed consistently well, Subtask 1L was the only one significantly correlated with fire support effectiveness.

The four subtasks rated by the fire support controller were:

1I. Plan use of organic/attached and non-organic fires.

1J. Determine priority of fires.

1L. Conduct initial fire support coordination.

7A. Modify fire support plan.

TABLE 15
Number of Fire Support Items Significantly Correlated ($p < .01$) With Effectiveness Ratings

Effectiveness Rating	Mechanized CFO	Mechanized Attack	Non-Mechanized Defense	Non-Mechanized Attack
FS	5	6	6	1
CG	0	0	0	0

NOTE: The total number of items was 14 in the covering force and defense, 13 in the attack.

TABLE 16
Number of Items Rated by the Monitor Significantly Correlated ($p < .01$) With Effectiveness Ratings

Effectiveness Rating	Mechanized CFO	Mechanized Attack	Non-Mechanized Defense	Non-Mechanized Attack
Msn	0	0	3	0
Cdr	2	0	5	1
CG	0	0	3	0

NOTE: The total number of items in each mission was 47.

Monitor. Few of the items rated by the monitor were correlated with any measure of overall effectiveness (Table 16), probably because each command group was observed by a different monitor.

Summary. Figure 4 presents the percentage of items on each observation form that were significantly correlated with one or more measures of effectiveness. In the two mechanized missions and in the defense, many of the S1/S4 items (39% to 65%) and the fire support items (43% to 46% were correlated with effectiveness. Few of them (14% and 8%) were related to effectiveness in the non-mechanized attack, when they were generally performed well. About one-fourth of the S2/S3 items (22% to 29%) were strongly related to effectiveness in all four missions. Few of the monitor's ratings (0% to 13%) were correlated with effectiveness in any mission.

Comparison with Figure 3 shows that many more items were correlated with effectiveness than were deficient. This was especially true in the defense, where a substantial percentage of the items were significantly related to effectiveness, but few of them were deficient.

CRITICAL PERFORMANCES

In terms of the criteria employed in this investigation, the most important command group performances were those that were both deficient and significantly correlated with overall effectiveness. Twenty-four of the performances evaluated in the areas of administration and logistics, intelligence and operations, and fire support were identified as critical in at least one mission. These performances and the missions in which they were critical are listed in Tables 17 through 19, below.

Administration and Logistics. Four items evaluated by the S1/S4 controller were both deficient and correlated with effectiveness. Three of them were critical in the covering force operation and one in the defense, as shown in Table 17. They are labelled in the table as they were on the controller's observation form. The first three items were elements of Subtask 9A, Arm and fuel the systems. The fourth item was Subtask 9D, Integrate CSS into the scheme of maneuver.

Intelligence and Operations. The 17 items that were identified as critical in this area are listed in Table 18. They are numbered consecutively in each category. They were renumbered in the table, because they did not have the same numbers on the observation forms for each mission. Five of the items corresponded to ARTEP subtasks. These were items A2, A4, A7, D1, and I3 which referred to the following subtasks:

1B, 2A. Identify critical combat information and intelligence.

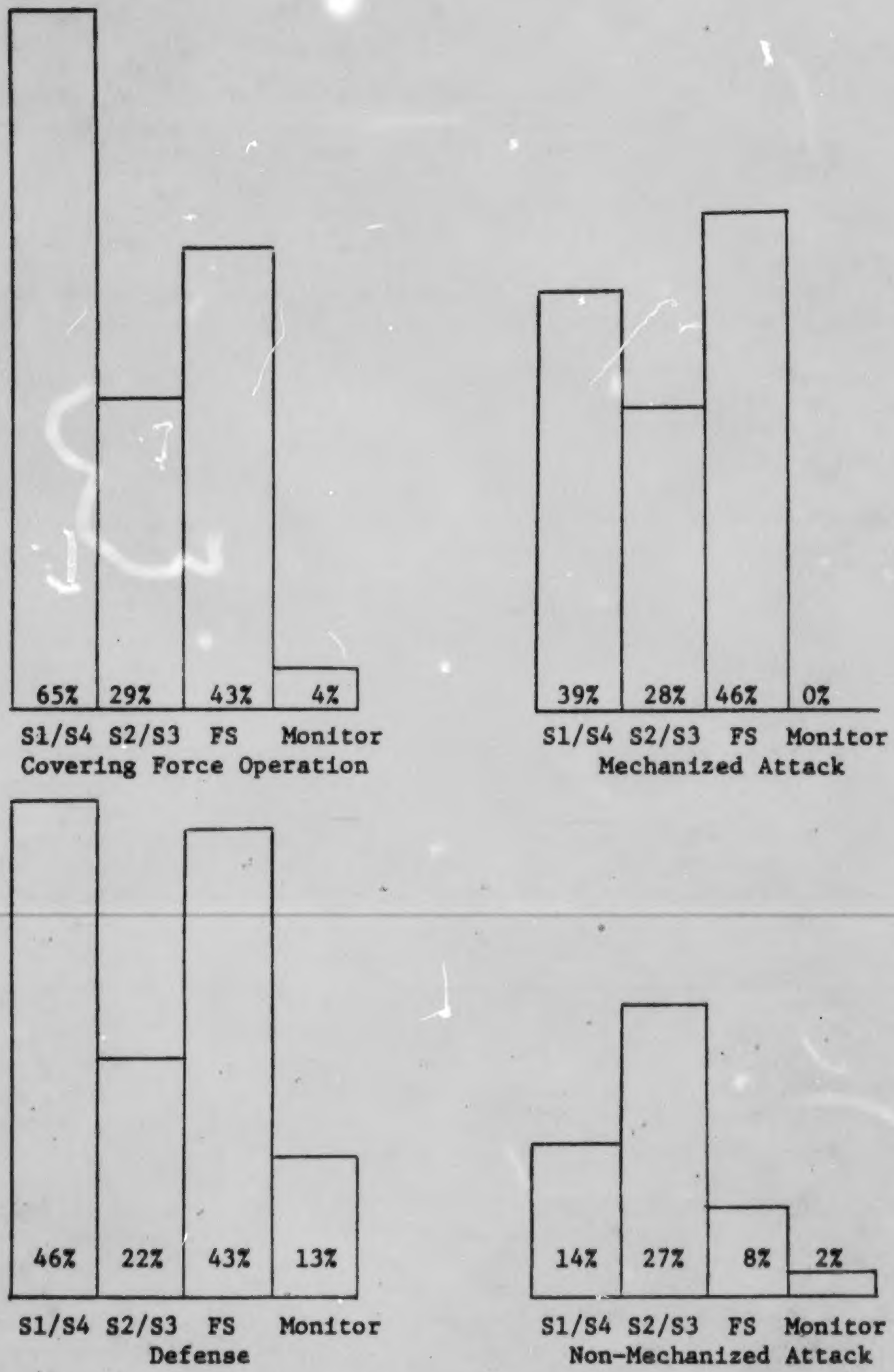


Figure 4. Percentage of items significantly correlated ($p < .01$) with effectiveness for each rater and type of mission.

TABLE 17
Critical Performances in Administration and Logistics

Item	Mission
6. In providing supplies and equipment to arm and fuel the system during the battle (9A):	
d. Did the S4 coordinate with the S2 so he knew the enemy's capabilities?	CFO
e. Did the S4 keep his higher appropriately informed of his activities?	CFO
g. Did the S4 effectively utilize his direct support assets?	Defense
9. c. How effectively was CSS integrated into the scheme of maneuver? (9D)	CFO

TABLE 18
Critical Performances in Intelligence and Operations

Item	Mission
A. Intelligence preparation of the battlefield.	
1. Was the enemy's scheme of maneuver and fire support identified?	CFO
2. Overall, how well did the command group identify critical combat information and intelligence? (1B, 2A)	CFO
3. Was the TF intelligence collection plan properly prepared, and did it reflect analysis by the battalion S2 of tasking responsibilities?	CFO, Non-mech atk
4. Overall, how well did the command group determine combat information and intelligence shortfalls and aggressively gather information from all available/appropriate sources? (2B)	CFO, Mech atk
5. Was relevant information from higher headquarters and adjacent units disseminated to company commanders (e.g., minefields)?	Mech atk
6. Were company commanders given an estimate of specifically what they would be facing?	CFO
7. Overall, did the command group disseminate combat information and intelligence that was event-oriented and usable to the recipient? (2D)	CFO
D. Communicate/coordinate plans and orders.	
1. Overall, were the orders appropriate, clear, concise, and did they contain essential information; were they issued so as to allow TF elements maximum time to go through troop leading procedures; and were they coordinated with proper agencies? (3G)	CFO, Mech atk
F. Troop lead during the battle.	
1. Were all attached combat units adequately controlled/monitored during the conduct of the exercise? (11A)	CFO

TABLE 18 (Continued)
Critical Performances
in Intelligence and Operations

Item	Mission
G. Coordinate/communicate changes.	
1. Did the command group sometimes assume all commanders were monitoring radios for changes? (6B)	Mech atk
H. Concentrate/shift combat power.	
1. Were tactical decisions made consistent with the time-distance relationship? (8C)	CFO
I. Enemy EW considerations.	
1. Was there too much communication?	Mech atk
2. Did security violations occur during radio traffic?	CFO, Mech atk
3. Overall, how well did the command group adhere to communications and electronic security measures? (10A)	CFO, Mech atk
4. Did the command group direct a switch to spare frequency as a last resort, using proper authentication techniques? (12A)	CFO
J. Other.	
1. Was there sufficient intra-staff coordination between 2/3 and 1/4?	Mech atk
2. How well did the command group apply the time-distance relationship while maneuvering Task Force elements?	CFO

2B. Gather critical combat information and intelligence.

2D. Disseminate critical combat information and intelligence.

3G. Communicate/coordinate plans and orders.

10A. Defeat or suppress opposing force's electromagnetic intelligence effort.

The subtasks to which the above items referred are indicated in parenthesis after the items in the table. Six other items, A1, A3, A5, A6, I1 and I2, were elements of these subtasks. Four more items, F1, G1, H1 and I4 were elements of the following subtasks:

11A. Troop lead during battle.

6B. Coordinate/communicate changes.

8C. Concentrate/shift combat power in the defense or retrograde.

12A. React to opposing force electronic warfare.

Finally, items J1 and J2 were not classified as part of any specific subtask. Four of the items in Table 18 were critical in both the covering force and the mechanized attack, one in the covering force and nonmechanized attack, eight in the covering force alone, and four only in the mechanized attack. None were critical in the defense.

Fire Support. The three items in Table 19 are labelled as they were on the fire support observation forms. Item 1a was part of Subtask 1I, Plan use of organic/attached and non-organic fires. Item 4b was part of Subtask 7A, Modify fire support plan, and Item 4c referred to the entire Subtask 7A. All three performances were critical in the covering force.

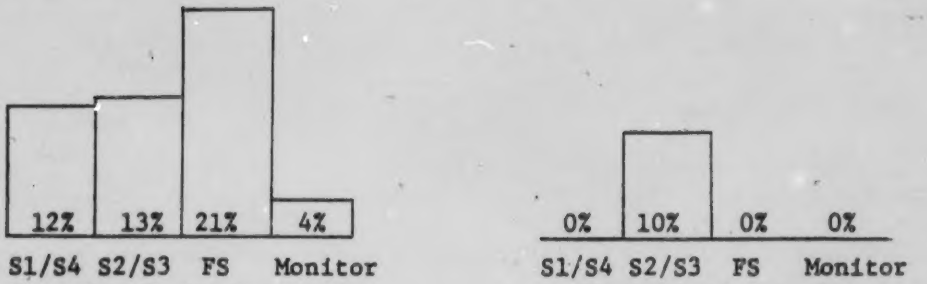
Monitor. Two items rated by the monitor met the joint criteria of deficiency and correlation with effectiveness: Subtask 8A, Determine critical place and time, and Subtask 8C, Concentrate/shift combat power in the defense or retrograde, were both identified as critical in the covering force operation.

Summary. It is apparent in Figure 5 that the majority of critical performances were identified in the covering force operation, where 12% to 21% of the S1/S4, S2/S3 and fire support items were both deficient and significantly correlated with effectiveness. A secondary concentration of critical performances (10%) occurred in the S2/S3

1306

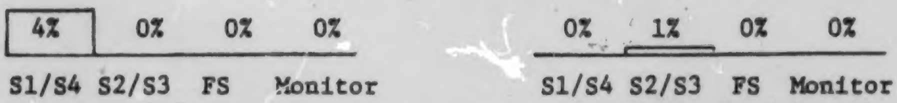
TABLE 19
Critical Performances in Fire Support

Item	Mission
1. Plan use of organic/attached and non-organic fires (1I)	
a. Did fire plan effectively utilize heavy mortars?	CFO
4. Modify fire support plan. (7A)	
b. Were requests for immediate fire support received and assigned to appropriate fire support agencies?	CFO
c. Overall, how well did the command group perform relative to the standard?	CFO



Covering Force Operation

Mechanized Attack



Defense

Non-Mechanized Attack

Figure 5. Percentage of items identified as critical by each rater for each type of mission.

area in the mechanized attack. Very few items were critical in the defense or the non-mechanized attack.

RATER RELIABILITY

Comparing the performance ratings from several different observers of the same command group, suggested three related questions:

1. How reliable were the ratings of a single observer?
2. How reliable were the mean ratings from several observers?
3. How significant were the differences among raters?

Since the intelligence and operations items were scored by four or five raters at every exercise, it was possible to measure the amount of consistency or disagreement among different raters. Rater reliability measures the consistency within one or more raters; conversely, analysis of variance tests the significance of differences among raters.

Reliability is defined as the ratio of true score variance to total score variance, where "true score" means that part of the score that is the same at each rescoring. Two measures of rater reliability[9] are the reliability of ratings from a single rater, r_{11}, and the reliability of mean ratings from k raters, r_{kk}: $r_{11} = \frac{Vi - Ve}{Vi + (k-1)\ Ve}$ and $r_{kk} = \frac{Vi - Ve}{Vi}$

where Vi = variance for items
Ve = variance for error
k = number of raters

The coefficients of rater reliability for eight randomly selected missions are listed in Table 20 in the order that the missions occurred. Each coefficient was calculated from the intelligence and operations items rated on a five-point scale by four or five observers. The reliability of ratings from a single rater varied from .07 to .38 with a mean of .22. The reliability of mean ratings from several observers varied from .29 to .71 with a mean of .55. Thus, increasing the number of raters from one to four or five more than doubled the rater reliability.

The variance analyses on which the reliability estimates were based showed that in every case the differences among raters were statistically

---

[9] Guilford, J. P. Psychometric Methods, 2d Ed. New York: McGraw Hill, 1954.

TABLE 20
Rater Reliability
for Eight Selected Missions

Number Raters, k	Coefficient of Reliability Single Rater, r_{11}	Mean of k Raters, r_{kk}
4	.24	.56
5	.21	.57
4	.21	.52
5	.26	.64
4	.38	.71
5	.24	.61
5	.18	.53
5	.07	.29

significant beyond the .001 level. In a typical case, with $r_{11} = .21$ and $r_{55} = .57$, the mean ratings from five different observers, averaged over all the five-point intelligence and operations items, were 2.6, 3.5, 3.8, 3.9 and 4.4. Significant differences among the means of different raters for the same command group is evidence of rater bias, i.e., the tendency of an individual to give high, or low, ratings. One implication of this result is that the same rater or, preferably, raters should be employed when comparing the performances of different command groups or of the same command group at different times.

DISCUSSION

CRITICAL SUBTASKS

The present investigation brings to 50 the total number of battalion command groups whose performance in CATTS exercises has been analyzed. Twenty-three groups were observed in the present study and 27 in the previous one.[10] Both investigations used essentially the same criteria of low performance rating and high correlation with effectiveness to identify critical performances, although there were some differences in detail. One difference was that in the present investigation a subtask was considered critical when any one of its elements was critical; whereas in the previous investigation subtask elements were not rated. Another difference was in the cut-off point for identifying low-rated performances. In the present investigation, a performance was classified as deficient when its mean rating was at or below the midpoint of the rating scale. In the previous investigation, a performance was considered deficient when its mean rating was one standard deviation below the mean of all the subtasks rated by the observer who evaluated it.

The reason a relative criterion was used in the previous investigation, rather than the midpoint of the scale, was that very few subtasks were rated below the middle of the three-point scale used in that study. The three points were defined as:

1 - Unsatisfactory, major departure from ARTEP standard

2 - Minor deviation from standard

3 - Satisfies the standard,

---

[10] Barber, H F., and Kaplan, I. T. op. cit.

and the lowest rating was rarely given. In contrast, the five-point scale used in the present investigation increased the range of scores given for subtask performance, so that for every rater there were several subtasks whose mean ratings were below the midpoint of the scale.

In spite of some differences in procedure between the two investigations, there was considerable agreement between them with respect to the subtasks that were identified as critical. Table 21 summarizes all the subtasks that were critical in either study - a total of 20 different subtasks. The X's in the two right-hand columns indicate the study in which each subtask was critical. It can be seen that 14 subtasks were identified in the previous investigation and 15 in the present one. Nine subtasks were critical in both studies: Subtasks 1B, 1I, 2A, 2B, 3G, 8A, 8C, 10A and 12A.

Furthermore, all the subtasks that were critical in only one study met one criterion of criticality in the other. Thus, five subtasks (2C, 5B, 5C, 5D and 8B) were critical in the previous investigation, but not in the present one. Four of them (2C, 5B, 5C and 5D) were deficient in the present investigation; however, they were not correlated with effectiveness. The fifth (8B) was correlated with effectiveness, but was not deficient. Conversely, six subtasks (2D, 6B, 7A, 9A, 9D and 11A) that were critical in the present investigation were not critical in the previous one. Subtask 2D was deficient in the previous study, but it was not then correlated with effectiveness. The other five subtasks (6B, 7A, 9A, 9D and 11A) were correlated with effectiveness, but were not deficient. Overall, the two investigations of battalion command group performance were in general agreement concerning the identification of critical ARTEP subtasks.

The four missions, however, differed greatly with respect to subtask criticality. All but one (6B) of the 15 subtasks that were critical in this investigation were critical in the covering force operation. Five subtasks (2B, 2D, 3G, 6B and 10A) were critical in the mechanized attack. Only one (9A) was critical in the defense, and one (2B) in the non-mechanized attack. The distribution of critical subtasks paralleled the distribution of performance deficiencies, i.e., the greatest number of deficiencies occurred in the covering force mission followed by the mechanized attack, defense, and non-mechanized attack.

Performance differences between the first and second missions may have been caused by improvement with practice or by differences in difficulty. The contributions of practice and difficulty were confounded, because the second mission was always an attack. The results do indicate, however, that the mechanized missions were more difficult

TABLE 21
Subtasks Identified as Critical for Battalion Command Group Training

		Previous Study	Present Study
TASK 1.	Develop plan based on mission.		
1B.	Identify critical combat information and intelligence.	X	X
1I.	Plan fires.	X	X
TASK 2.	Initiate intelligence preparation of the battlefield.		
2A.	Identify critical combat information and intelligence.	X	X
2B.	Gather critical combat information and intelligence.	X	X
2C.	Analyze opposing force.	X	
2D.	Disseminate critical combat information and intelligence.		X
TASK 3.	Prepare and organize the battlefield.		
3G.	Communicate/coordinate plans and orders.	X	X
TASK 5.	See the battlefield during the battle.		
5B.	Gather critical combat information and intelligence.	X	
5C.	Analyze opposing force.	X	
5D.	Disseminate critical combat information and intelligence.	X	
TASK 6.	Control and coordinate combat operations.		
6B.	Coordinate/communicate changes.		X
TASK 7.	Employ fires and other combat support assets.		
7A.	Modify fire support plan.		X
TASK 8.	Concentrate/shift combat power.		
8A.	Determine critical place and time.	X	X
8B.	Concentrate/shift combat power in the attack.	X	
8C.	Concentrate/shift combat power in the defense or retrograde.	X	X
TASK 9.	Manage combat service support assets.		
9A.	Arm and fuel the systems.		X
9D.	Integrate CSS into scheme of maneuver		X
TASK 10.	Secure and protect the TF.		
10A.	Defeat or suppress opposing force's electromagnetic intelligence effort.	X	X
TASK 11.	Troop lead during battle.		
11A.	Supervice compliance with TF order.		X
TASK 12.	React to situations requiring special actions.		
12A.	React to opposing force electronic warfare.	X	X

than the non-mechanized ones. In the previous investigation, there were too few non-mechanized units to permit a similar comparison of mission performance.

RATER RELIABILITY

The utility of the command group ARTEP as a means of diagnosing training deficiencies and evaluating command group effectiveness is limited by its reliability as a measuring instrument. Even under the ideal conditions of this investigation (i.e., experienced controller-evaluators, a realistic, automated battle simulation, and a standardized scenario) rater reliability was low. Low reliability can be tolerated in research, when ratings from many units can be averaged so that rating errors tend to cancel out. However, low reliability is a problem when performance ratings are used to diagnose and evaluate individual command groups.

Two steps can be taken in the present time frame to enhance the reliability of command group performance ratings: one is to average the ratings from several different observers; the other is to use the same raters when comparing different command groups or when evaluating the performance of a given group at different times. Over the longer term, however, the improvement of rater reliability will require continued research to develop more objective measures of command group performance.

SUMMARY AND CONCLUSIONS

1. Fifteen subtasks of the command group/staff module of ARTEP 71-2 were identified as critical by virtue of being both low-rated and highly correlated with effectiveness. These subtasks can be summarized briefly within five functional areas:

a. Fire support: Develop (1I) and modify (7A) the fire support plan.

b. Intelligence preparation of the battlefield: Identify (1B, 2A), gather (2B) and disseminate (2D) critical combat information and intelligence.

c. Operations: Communicate/coordinate plans and orders (3G) and changes (6B), and supervise compliance with the task force order (11A). Determine the critical place and time (8A), and concentrate/shift combat power (8C).

d. Logistics: Arm and fuel the systems (9A), and integrate combat service support into the scheme of maneuver (9D).

e. Electronic warfare: Combat enemy electromagnetic intelligence (10A) and electronic warfare (12A).

These critical performances should be given particular attention in the development and evaluation of command group training programs and simulations.

2. The four missions observed in this investigation were markedly different with respect to subtask criticality. All but one (6B) of the subtasks listed above were critical in the covering force operation, five (2B, 2D, 3G, 6B and 10A) were critical in the mechanized attack, one (9A) in the defense, and one (2B) in the non-mechanized attack. The effects of practice and difficulty were confounded in comparing the attack missions to the covering force or defense, because the attack was always the second mission. It can be inferred, however, that the covering force was more difficult than the defense, since both those missions were performed first - by mechanized and non-mechanized groups, respectively.

3. Rater reliability, which measures the internal consistency of a rater or group of raters, was low. The coefficient of reliability for subtask performance scores from a single rater was only .22. It increased to .55 when the scores from four or five raters were averaged.

4. Individual raters differed in their judgement of subtask performance. The differences among ratings of the same command group by different observers were significant beyond the .001 level.

5. The effects of mission type, rater reliability, and individual differences among raters have implications for the measurement of command group performance. These effects should be controlled when diagnosing training requirements, comparing command groups, or evaluating training systems. Specifically, the same type of mission and the same raters (several raters) should be used when comparing the performance of different command groups or of the same command group at different times. In addition, the low rater reliability and significant differences among raters indicate the desirability of further research to develop more objective measures of command group performance.

AN APPLICATION OF TACTICAL ENGAGEMENT SIMULATION FOR UNIT PROFICIENCY MEASUREMENT

C. Mazie Knerr and Robert T. Root
US Army Research Institute for the Behavioral
and Social Sciences

LTC Larry E. Word
US Army Training Support Center

The need for methods of measuring team and unit proficiency, and the lack of knowledge in this area, are widely recognized. Team performance measurement difficulties have been noted as fundamental problems in unit proficiency diagnosis and training evaluation, in both military and civilian settings (Blum and Naylor, 1968; Defense Science Board, 1975). Existing combat unit performance measurement techniques depend largely on judgmental data, and often do not evaluate the unit's ability in the field (Hayes et al, 1977). Researchers must solve these measurement problems before they can substantially improve unit training.

A tactical training system, called engagement simulation (ES), that includes objective, accurate casualty assessment, offers a potential solution for team performance measurement in combat training. Objective casualty assessment provides the primary measures of team proficiency, such as casualty exchange ratios and mission accomplishment. Recent advances in ES procedures have further improved its use for assessing tactical performance. This paper reviews application of ES to unit measurement, with emphasis on lessons learned while validating ES procedures for armor units, and while developing ES for armored cavalry units.

ENGAGEMENT SIMULATION

ES techniques provide realistic tactical training under conditions that simulate the complex modern battlefield. In addition to the casualty assessment, characteristics of ES that contribute to the realism are that it uses two-sided, free-play tactical field exercises, and it simulates weapons effects and signatures.

Objective casualty assessment is achieved when a soldier, looking through a 6-power telescope mounted on his rifle, correctly reads a 3-inch, two-digit number on the helmet of an opposing unit member. The telescope power and helmet number size are calibrated to produce hit/kill

---

The views, opinions, and findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

probabilities realistic for the weapon's lethality. When the soldier fires a blank round and correctly identifies the opposing helmet number, a casualty is assessed. A controller with the fire team radios the helmet number to the controller with the opposing team, who informs the "hit" soldier (US Army Infantry School, 1975).

Analogous objective casualty assessment, weapons effects, and signature simulation procedures have been established for infantry, armor, and antiarmor elements, including these weapons systems: M60 tank main gun, mines, hand grenades, machine guns, and light (LAW), medium (DRAGON), and heavy (TOW) antitank weapons. For weapons with longer ranges than that of the rifle, the controller is equipped with optics to sight individual helmet numbers and numbers on panels attached to vehicles. In the tank, for example, the controller's telescope is mounted in the breech of the main gun. When the controller in the tank determines that the gun is centered on a target at the time of simulated round impact, he assesses a casualty. The controller then radios the number of the "hit" in the same way as described for the rifle casualty assessment.

The radio net over which the controllers announce the casualties is used by senior controllers to administer the exercise, and is monitored by personnel who record the "hits." They write the time, target number, and firer number on a net control sheet, and they check that the "hit" was confirmed by the controller in the target vehicle.

All ES systems provide some way of identifying casualties. In the REALTRAIN system, telescopes and numbers are employed, and have been used for training with opposing forces as large as reinforced platoons. In order to achieve tactical realism in larger units, a Multiple Integrated Laser Engagement System (MILES) has been developed. MILES employs low-power, eye-safe laser transmitters mounted on each weapon. Each target (vehicle or soldier) has solar cell detectors which receive the laser signal as either a hit or a near miss. Hits activate a buzzer on the target which can be silenced by deactivating the targets laser transmitter. The lasers are pulse coded to differentiate weapons' effects (e.g., rifles can kill individuals but cannot kill tanks). Employment of the lasers is expected to reduce the need for human controllers, and simultaneously, to reduce the amount of data on tactical activities.

ES differs from some of the more frequently encountered simulation techniques. It is not a board game or computer simulation, but is conducted in training fields, with a full complement of soldiers and equipment. Although it employs the tactical equipment, it emphasizes the human behavior: it is man-ascendant rather than machine-ascendant. The decisions, reactions to events that emerge during competition with a motivated, intelligent adversary, and other responses to the environment are emphasized. The cues to which soldiers must respond are the same as those to which they respond in battle, and the situation changes as a result of their actions. Thus, the situation is emergent rather than prespecified, highly predictable, or amenable to analytic solution (Boguslaw and Porter, 1966).

PERFORMANCE ASSESSMENT

The objective casualty assessment in ES provides some, but not all, of the necessary performance measurement. While casualties (target, firer, and time) are the primary criteria, relying solely on them makes it difficult to determine why they occurred. Additional observations, or measures of active performance, are required when the final outcome is not an adequate index of the skill (Cronbach, 1960). Measures of processes, or intermediate task and training objective performance, assist in training diagnosis and explanation of product data. An example is the detection and engagement of the enemy at the maximum possible range during defensive missions. Particularly at company level and below, there is little recognition of the importance of observation posts to provide exact and timely information concerning the enemy. In exercises between relatively untrained units, most critical decisions and actions occur along the forward edge of the battle area. As the units become more sophisticated, leaders in the defensive unit spend a greater percent of their efforts selecting observation post positions, planning communications and indirect fire, and positioning long range direct fire weapons. Consequently, detection and effective engagement ranges increase.

Tactical outcomes depend upon several factors other than the proficiency of the units: forces, missions, weather, and interactions among these factors can influence the tactical results. For example, weather interacts with force mixes, since poor visibility favors dismounted troops, to the disadvantage of long range weapons. If visibility improves during the tactical action, then the advantage reverts to the long range weapons. Due to these interactions, the outcome does not necessarily indicate the relative proficiency of the opposing forces. The impact of external factors must be considered before the results of an exercise can be used to diagnose proficiency.

Problems arise in both the recording of behavior (active performance, or processes) and the encoding of the environment (such as the external factors). Thus, observational field research needs a system for detecting, measuring, and recording the events and other factors pertinent to the action (Sells, 1966).

Literature on ratings and observational performance assessment techniques in criterion development offers suggestions to improve field measurement (Blum and Naylor, 1968; Goldstein, 1974; Guilford, 1954; Simon, 1969; Wherry, 1952). Observations and ratings of behavior can suffer from unreliability and inaccuracy due to a variety of error sources. First, the performance itself is inconsistent, since people perform better at some times and under some conditions than others. This is especially true in emergent situations, where a given behavior may not be required in a specific instance, or may be altered to suit the situation. Second, the detection, or observation of the behavior is unreliable. An observer may or may not detect a given behavior, and different observers may vary in correctness of perceiving and assessing it. Third, recording of behavior introduces error, depending on the type of record. For example, immediate recording of events as they occur reduces error by decreasing recall, or memory effects. Despite these error sources, observations and other judgmental measures continue to be the most frequently used type for performance criteria (Blum and Naylor, 1968).

Improved measurement can be achieved when the researcher (a) specifies and defines as concretely as possible the behaviors to be observed, (b) requires data collection personnel to observe, but not judge the behavior, (c) trains the observers fully, and (d) requires observers to record their observations immediately. The following sections discuss how we applied these principles, and used observational techniques in conjunction with objective measures.

OBJECTIVE MEASURES

The use of casualty, time, and mission accomplishment data is demonstrated by results from the validation of armor REALTRAIN (Scott, Meliza, Hardy, Banks, and Word, 1978). Teams composed of tanks, heavy antitank weapons (TOW), and artillery forward observers were pretested against a similarly composed opposition force (OPFOR). Half of the tested teams then had a week of REALTRAIN training, while the others had conventional tactical field training. The teams were posttested against the OPFOR. Casualty data show that the teams were similar in pretest performance (each bar in Figure 1 represents 52 vehicles). REALTRAIN teams improved in terms of casualties inflicted on the OPFOR, as seen in the posttest data, while conventionally trained teams did not.

Temporal distributions of the casualties during an exercise provide additional insight into changes in tactical performance. When the cumulative percent of tested unit casualties are plotted against the time lapsed, it appears that fewer casualties are sustained early in the exercises after REALTRAIN training, in contrast to heavy early losses before training (Figure 2). Conventionally trained units sustained heavy early casualties both before and after training. Time data, in association with other objective data such as casualties, can be of help in measuring what went on during an exercise and what may have led to successful (or unsuccessful) mission accomplishment.

Mission accomplishment results show the same patterns of REALTRAIN effectiveness as did the casualties. In order for a tested unit to accomplish its attack mission, it had to clear an objective of OPFOR elements and occupy the objective. To accomplish its defense mission, it had to prevent the OPFOR from occupying an objective for sixty minutes. Mission accomplishment data for both attack and defense missions are combined in Figure 3, where each bar represents 8 exercises. REALTRAIN teams improved in their ability to accomplish their mission successfully, while conventionally trained teams did not.

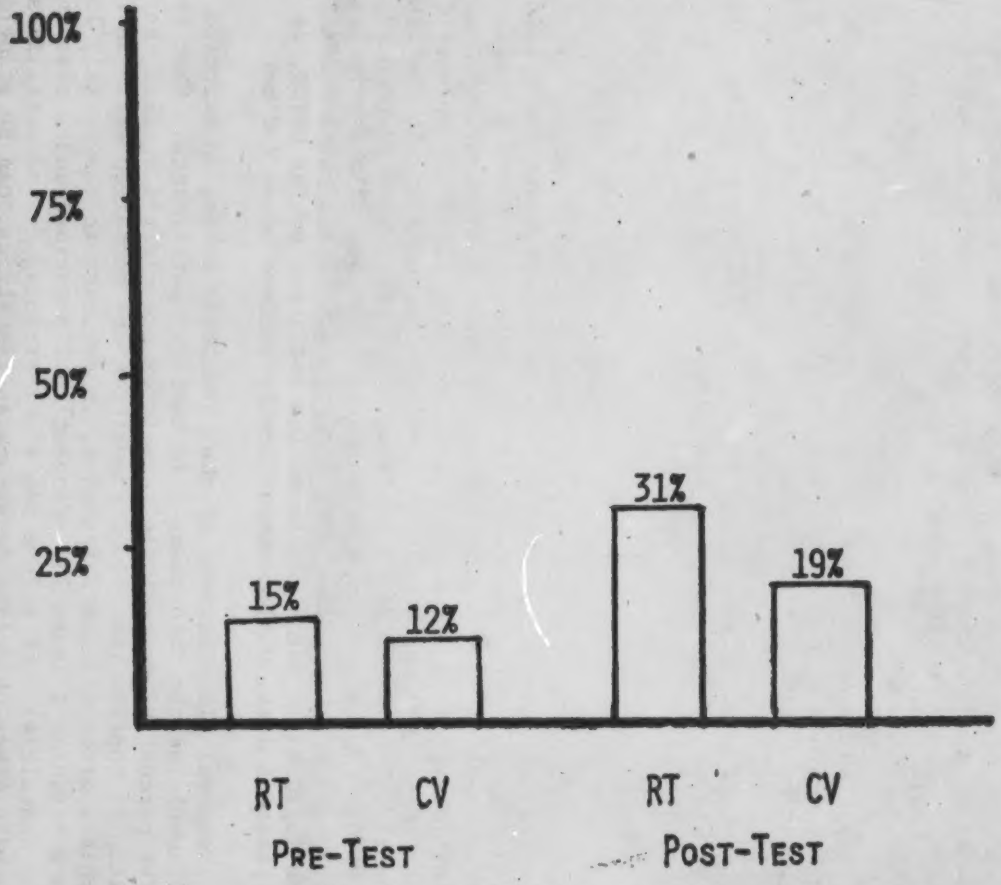
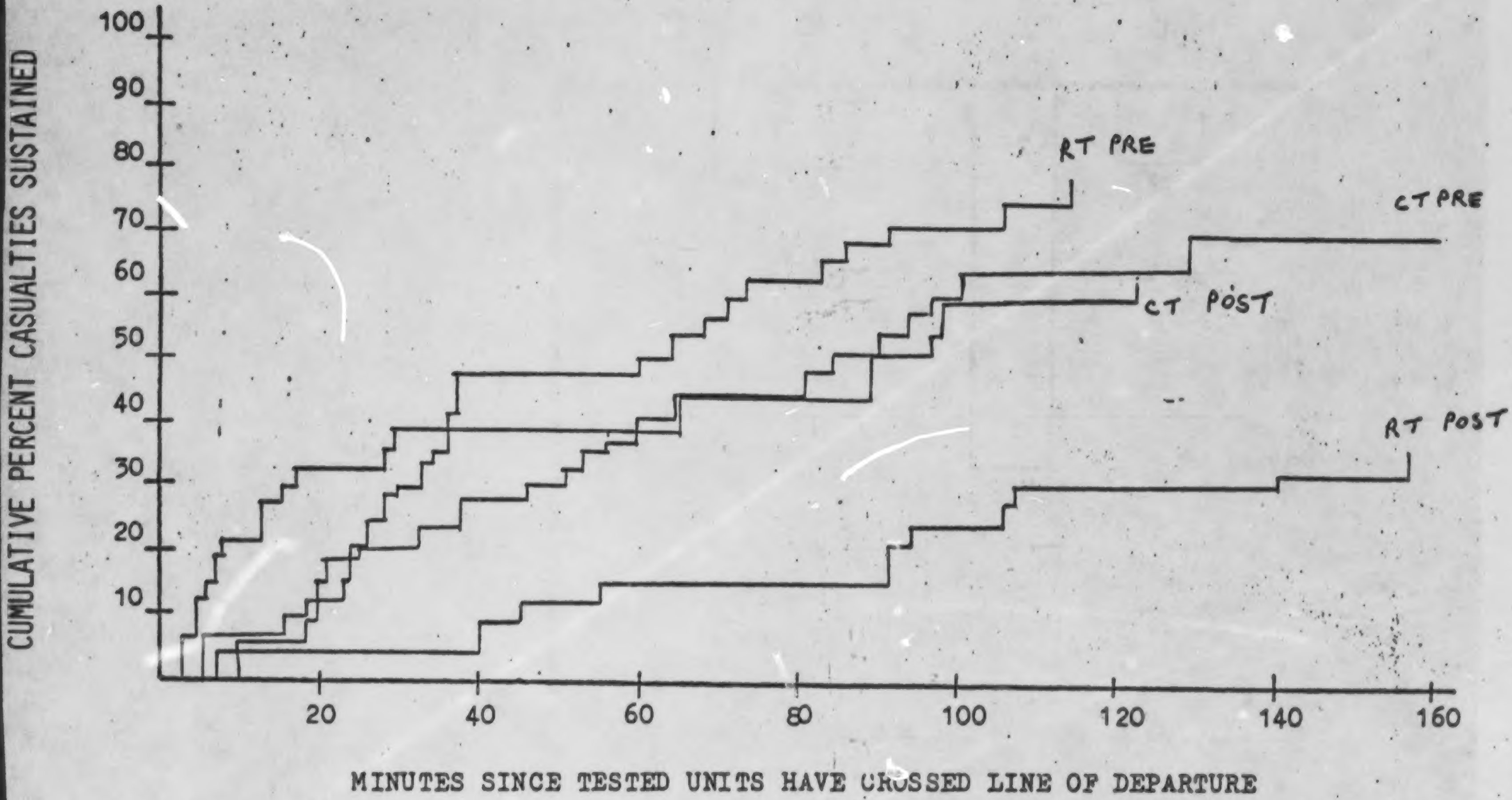


Figure 1. PERCENT OF OPFOR CASUALTIES: ARMOR TEST



1381 Figure 2. Effects of training and test on destruction of tested unit weapon systems as a function of time in the attack. 1382

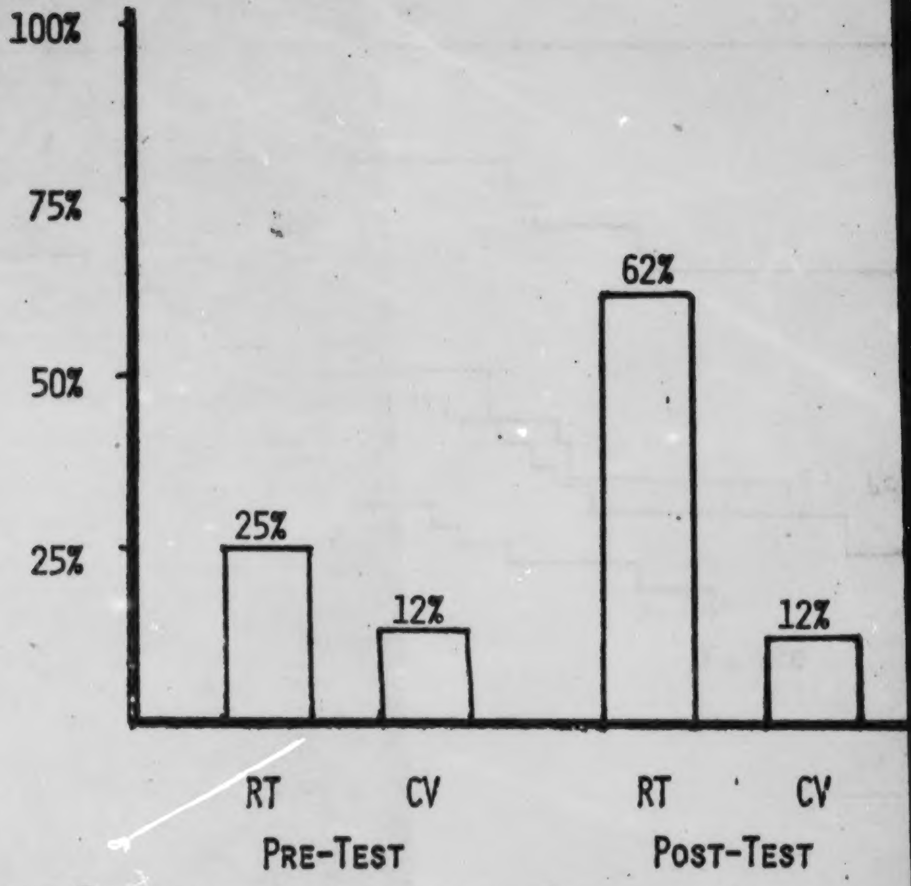


Figure 3. PERCENT OF SUCCESSFUL MISSIONS: ARMOR TEST

Other objective data, such as artillery fire planning and use, are also recorded. An indirect fire data form is completed by personnel in the fire direction center, indicating the number of rounds fired, time distribution, and casualties inflicted. The example in Figure 4 shows that "jeep 28" was hit by six rounds early in the exercise, but that no other indirect fire missions for this team were effective in this exercise. Inclusion of these data further clarifies explanation of the overall results.

ARMORED CAVALRY ENGAGEMENT SIMULATION

Unlike other ES applications, armored cavalry ES cannot rely on casualties as performance data. "Cavalry's basic tasks are reconnaissance and security" (Department of the Army, 1977), and may not be casualty-producing. The armored cavalry platoon performs information gathering and reporting functions. When reconnaissance units withhold fire (e.g., to avoid disclosing their positions), tactical events may not lead to casualties. While developing ES procedures for cavalry units, the problem was to develop a realistic training environment for the reconnaissance functions, while maintaining the objectivity and credibility of the casualty-producing ES exercises. Thus, the cavalry ES research focused on process measures and external factors.

An armored cavalry ES training program was designed with help from the training personnel from the unit providing support in the 3d Armored Cavalry Regiment, Fort Bliss, Tex. Research results have been reported previously (Knerr, Hamill, and Severino, 1978; Knerr, Stein, Hamill, and Severino, 1978).

Only two weeks were available for the program, so that it was not feasible to test all combinations of missions, force structures and force ratios. The armored cavalry force was a regimental cavalry platoon, containing scout, light armor, infantry, and mortar sections. The OPFOR was a combined arms team composed of tank, TOW, and infantry sections, with simulated indirect fire support. For each mission, the OPFOR composition was varied to enhance realism and provide reasonable opposition. The missions selected were reconnaissance (area, route, and zone) and delay (Table 1).

In these exercises, weather and terrain had strong effects on mission accomplishment. The weather was clear and sunny, providing optimal visibility. The terrain was flat desert, although there were sand dunes that could hide vehicles and soldiers. Moving vehicles were quickly detected by exhaust smoke and dust clouds from the tracks. The force assigned an attack mission, or moving mission of any sort, was at a disadvantage under these conditions.

Relative combat power interacted with other external factors. Results of an attack with a 3 to 1 force ratio differ from results with 6 to 1 odds. If the opposing force is either too strong or too weak,

PREPLANNED FIRES

Trial No. CA#4
Date 24 MARCh 78
Unit I.D. Team A
Name Sp/4 DOUGlas

Test X
Training
Shoot-Off
CAT

101 1. 092723
102 2. 074704
3.
4.
5.
6.
7.
8.

Yellow Request Time	Requesting I.D. (KT No.)	Location/ Adjustment	O-T Azimuth	Splash Time	RDS Fired	Registration	Suppression	Smoke	IE	Casualties
10:08	B1. 47	092723		006 007 10:14	6sim				—	JEEP 28 DESTROYED 4 PERSONAL EXPOSED. TK 13 COMMO. 2 PERSONAL EXPOSED
10:16		-100 FFE	3200	012 10:20	6sim					No cas.
10:22				END miss	NONE				—	—
Yellow										
10:30	B 1. 47	091729		10:35 10:34	1sim 1sim	→ REPEAT → DUD			—	NO cas.
10:36		R200	3200	10:42 10:38	1sim 1sim	→ REPEAT			—	No cas.
10:44				END miss	NONE				—	—
Blue										
10:44	B1 47	074704 FFE		10:53	6 sim				—	No cas
10:54				END miss	NONE				—	—
A	B	C	C	E	F	G	H	I	J	K

FIGURE 4. INDIRECT FIRE DATA FORM

TABLE 1

PLATOON MISSIONS BY EXERCISE

EXERCISE	CAVALRY PLATOON MISSION	OPFOR PLATOON MISSION
1	ZONE RECONNAISSANCE	DELAY
2	ROUTE RECONNAISSANCE	SCREEN
3	FLANK GUARD	ROUTE RECONNAISSANCE
4	AREA RECONNAISSANCE	DELAY
5	ROUTE RECONNAISSANCE	ATTACK
6	DELAY/DEFEND IN SECTOR	ZONE RECONNAISSANCE

differences between the units may not emerge due to "ceiling" effects. During the first two days, the cavalry had reconnaissance missions and the OPFOR had a strong composition (main battle tanks, TOW, and infantry). After being hit hard on the first day, the cavalry moved so slowly on the second day that it made little progress. They did send reports of enemy strength to the commander, and it was realistic that they did not move forward in a "suicide" mission against the heavy, long range weapons they detected. On subsequent days, the OPFOR was reduced and the action was more realistic.

External factors (missions, terrain, weather, forces) need to be considered in interpreting mission outcomes as measures of unit proficiency in tactical situations. Figure 5 shows the outline of a data form used to describe the exercise. The record starts with a description of the exercise lane (usually augmented by a map or sketch), weather, general tactical situation, missions, force structures, and other external factors or chance events. Next are notes of the platoon leaders' plans, and orders to the vehicle commanders. Complete notes of the tactical activities are then recorded, along with the mission outcomes. These notes on plans, orders, and tactical activities provide an overview of processes, i.e., active performance during the exercise.

PROCESS MEASURES

Process measurement in the armored cavalry ES development was based on the principles described earlier for the improvement of observational measurement: train observers; specify the behavior to be observed as precisely as possible; and record during the action. Observers received initial training during three days of small scale exercises that preceded the full scale platoon exercises. These small exercises also familiarized the observers with the terrain, equipment, maneuvers, and data collection forms. Observers were thoroughly briefed each day on the exercise scenario, operations orders, and anticipated tactical events.

In the first exercise, the cavalry platoon had a zone reconnaissance mission. To clarify the behavior to be observed and recorded, more detail was needed than is given in the cavalry ARTEP (Figure 6; Department of the Army, 1976). To perform effectively, the commander needs to know the location and status of friendly forces, and the location and strengths of enemy forces. The reconnaissance elements had to learn the importance of detecting the enemy at the maximum possible range, and reporting the information to the commander. For example, they had to provide exact and timely reports concerning the enemy to enable effective use of indirect fire.

To support these training objectives, the operations orders for the first exercise gave the cavalry platoon a zone reconnaissance mission, with the request that they provide early warning, occupy an objective by a given time, and prepare to defend. Specific elements of intelligence and coordinating instructions also clarified their assignment. Essential

LOCATION

TERRAIN DESCRIPTION

DATE EXERCISE NUMBER

GREEN TEAM
  ELEMENTS
  MISSION
  PLAN

BROWN TEAM
  ELEMENTS
  MISSION
  PLAN

OUTCOME

DISCUSSION

FIGURE 5. EXERCISE NARRATIVE

TRAINING AND EVALUATION

UNIT: Armored Cavalry Troop
MISSION: 1-7. Zone Reconnaissance

TASK	CONDITIONS	TRAINING/EVALUATION STANDARDS	RATING		
			S	U	REMARKS
		Elements conduct movement according to troop commander's task-organization.			
		Elements maintain OPSEC (see mission 0-15).			
1-7-6. Cross LD.	Troop commander designates the LD passage points, latest time to return through friendly lines, and other control measures in OPORD.	Elements cross LD on time.			
		Elements cross LD at designated passage points in specified task organization and begin zone recon.			
1-7-7. Reconnoiter designated zone.	Troop commander specifies task organization, command control, and boundaries in OPORD.	Recon elements report on Threat forces, key terrain and routes timely and within specified tolerances.			
		Elements conduct zone reconnaissance using proper movement techniques (FM 17-95).			
		Elements thoroughly search for Threat forces throughout zone.			
1-7-8. Make contact.	Threat force engages elements of troop.	Overwatch elements lay down suppressive fire and request indirect fire support.			
		Bounding elements deploy to cover and return fire.			
		Elements observe and report. Report includes type and number of vehicles in Threat force within 80% accuracy.			
		Elements in contact request permission to bypass Threat force. Detached element watches Threat force while remainder of troop continues zone reconnaissance.			

FIGURE 6. ARMY TRAINING AND EVALUATION PROGRAM FOR ARMORED CAVALRY
(ARTEP 1[illegible]55)

elements of intelligence included enemy left in the area, enemy strong points, and enemy ability to move forward. In the coordinating instructions, the unit was requested to hold at phase lines and request permission to cross, and to bypass pockets of resistance. They were under weapons hold status, in which they could fire only with permission of the commander. Thus, the general requirements in the ARTEP mission were stated more specifically, and observable activities were defined.

The general situation described in the operations orders was realistic for a weapons hold situation. As a result of this status, the vehicle commanders frequently reported enemy information, along with repeated requests for release from weapons hold and consequent permission to fire. They used their reports to build a convincing requirement to fire. The weapons hold status, applied in the highly motivating ES environment, appeared to elicit concentrated reconnaissance reporting.

Establishing the reporting requirements, and reinforcing them using the weapons hold status, made tactical communications a valuable data collection vehicle. The reports contained time and location information for both friendly and enemy elements. The quality of the data was a problem, in both accuracy and completeness due to radio problems, and reliance on tactical participants' own skills in location reporting. The controllers (e.g., troop commander) had to check the accuracy of the location information, but at least their task was narrowed to a manageable size by their having the other report information.

Additionally, the report data could be corroborated in many instances by its relation to objective data. In this first exercise, there were six casualties, providing objective information to confirm detections of enemy activity that were reported in the same time periods. Also, some conditions were established to create known situations, serving as probes to test reconnaissance capability.

In the second exercise, items of intelligence interest were placed at three known locations, as shown on the map sketch used to brief the observer (Figure 7): an abandoned armored personnel carrier, some weapons, and an enemy soldier (represented by a mannequin). Reports from one of the rifle squads early in the exercise indicated that the squad was not where it should have been; however, there was no way to be sure of their actual location. When they reached the abandoned vehicle and correctly reported its REALTRAIN number, however, it was certain that they had followed the wrong road. The mannequin "enemy soldier" also enabled observers to record the location of tactical events. Figure 8 shows a map record noted by an observer on the cavalry platoon leader's vehicle. The controller recorded the times and places that the platoon leader dismounted to conduct ground reconnaissance. These estimates were verified when the vehicle reached the known location of the mannequin and "took dummy prisoner at 1255". The observational data were thus anchored to a known location. In general, the known locations clarified records of tactical performance.

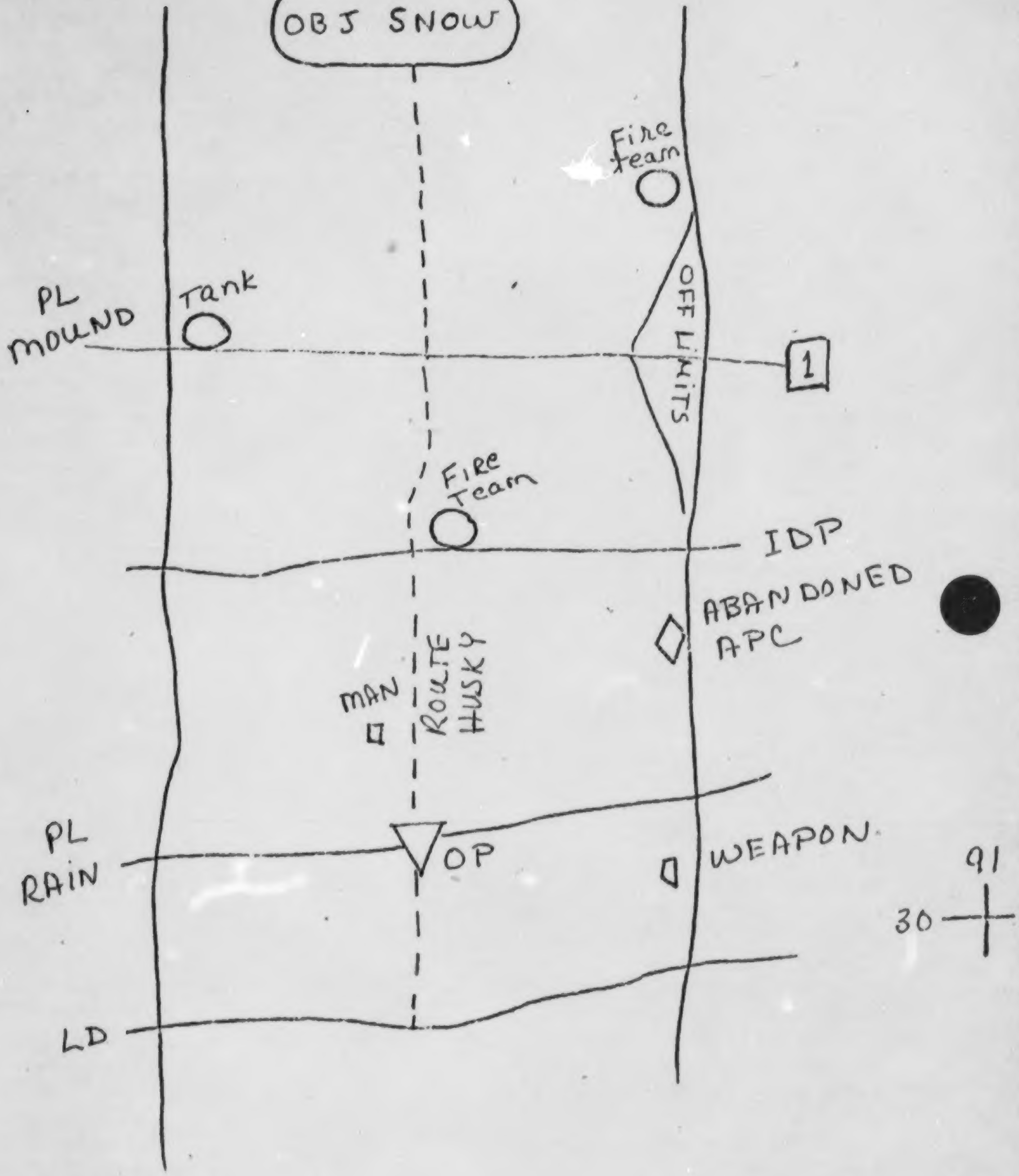

OBJ SNOW
Fire team
Tank
PL MOUND
OFF LIMITS
1
Fire Team
IDP
ABANDONED APC
MAN
ROUTE HUSKY
PL RAIN
OP
WEAPON
91
30
LD

FIGURE 7. MAP SKETCH

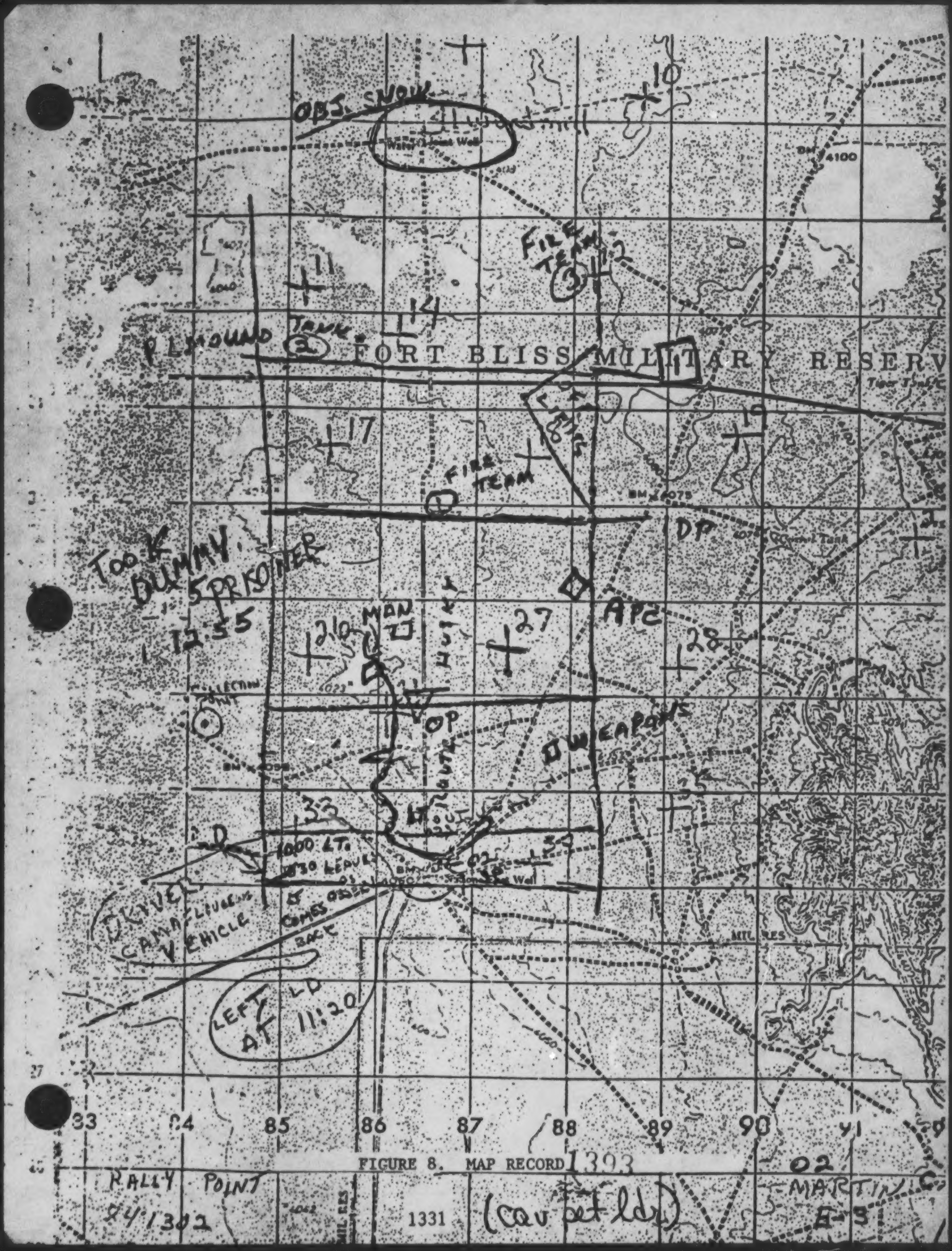

OBJ SNOW
Water Well
10
4100
FIRE TEAM 2
3
11
14
PLATOON TANK 3
FORT BLISS MILITARY RESERV
LIMITS
17
18
19
FIRE TEAM 1
BM 4075
DP
TOOK DUMMY PRISONER
5
12:55
MON
210
ROUTE
27
APC
28
COLLECTION POINT
OP
WEAPONS
BM 4035
33
1000 LT.
1030 REPUL
34
Well
DRIVE CAMOFLAGED VEHICLE
COMES BACK
LEFT LD AT 11:20
MIL RES
27
33
84
85
86
87
88
89
90
91
RALLY POINT
84 1302
1331
(cav set ld)
02
MARTIN
E-3

FIGURE 8. MAP RECORD 1393

DISCUSSION

Often in performance assessment situations, there is a strong tendency to measure what is easy to measure. For example, the Army Training Tests, which preceded the current ARTEPs, relied heavily on subjective checklists concerning the planning, coordination, preparation, and movement phases of tactical operations. ARTEPs emphasize the importance of analyzing critical aspects of missions. The major tasks differentiated for each mission in the ARTEP reflect fundamentals of land combat more accurately than did the earlier Army Training Tests. However, extensive training experience with tactical ES has demonstrated that further improvement can be made in the selection of training objectives and the measurement of the attainment of the objectives.

The application described in this paper started with the explicit definitions of some of the processes, or intermediate training objectives. The objectives were clarified by instructions in the operations orders that resulted in use of tactical reports to augment the data collection. The reports were corroborated using probes (of known location) and casualty reports. Thus, data of questionable accuracy were linked, where possible, with more accurate data. Observations were recorded during the exercises by data collectors who were thoroughly versed in the training objectives, tactical situation, missions, probes, and special techniques such as use of weapons hold.

This paper has focused on the nature of tactical data attainable using ES operations to acquire objective data, and methods of enhancing the accuracy of data. ARI is also working on the improvement of data collection and analysis, using an Automated Tactical Operations Measurement System (ATOMS), with contractual support from Human Sciences Integrated. ATOMS is comprised of data collection instruments, associated data collection and reduction procedures, and a software package for summary descriptive statistics from which further analyses may be made (Epstein, 1978; Root, Knerr, Severino, and Word, 1978).

The inherent difficulty of measuring complex human performance in a field environment accounts in part for the shortage of satisfactory methods for unit performance measurement (Wagner, Hibbits, Rosenblatt, and Schulz, 1977). The methodology described here depends on the whole system, from clarification of the training objectives, objective oberservation and recording, analysis, and explanation in sufficient detail to show how and why outcomes, such as mission accomplishment, occurred.

REFERENCES

Blum, Milton L., & Naylor, James C. Industrial Psychology. New York: Harper & Row, 1968.

Boguslaw, R. & Porter, E.H. Team functions and training. In R.M. Gagne and A.W. Melton (eds), Psychological Principles in System Development. New York: Holt, Rinehart, and Winston, 1966.

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper & Row, 1960.

Defense Science Board. Report of the Task Force on Training Technology, Chapter 8. Crew, Group, Unit Training, Washington, D.C.: Office of the Director of Defense Research and Engineering, Department of Defense, May 1975.

Department of the Army. Army Training and Evaluation Program for Armored Cavalry Squadron and Armored Cavalry Troop (ARTEP 17-55). Fort Knox, KY: US Army Armor School, June 1976.

Department of the Army. Field Manual No. 17-95, Cavalry, Washington, D.C., 1 July 1977.

Epstein, K.I. Automated Tactical Operations Measurement System - ATOMS. Paper presented at the Military Operations Research Society 41st Symposium, Washington, D.C. July 1978.

Goldstein, Irwin L. Training Program Development and Evaluation. Monterey, Calif.: Brooks/Cole, 1974.

Guilford, J.P. Psychometric Methods. New York: McGraw Hill, 1954.

Hayes, R.E., Davis, P.C., Hayes, J.J., Abolfathi, F., Harvey, B., & Keynon, G. Measurement of Combat Effectiveness in Marine Corps Infantry Battalions. Arlington, VA: Defense Advanced Research Projects Agency, October 1977.

Knerr, C.S., Hamill, B.W., & Severino, A.A. Engagement Simulation for Armored Cavalry: Initial Test. ARI Research Problem Review 78-5. August 1978.

Knerr, C.S., Stein, E.S., Hamill, B.W., & Severino, A.A. Development and Evaluation of Armored Cavalry Engagement Simulation. Paper presented at the Sixth Psychology in the DOD Symposium, US Air Force Academy, Colorado. April 1978.

Root, R.T., Knerr, C.S., Severino, A.A., and Word, L.E. Tactical engagement simulation training: A method for learning the realities of combat. Paper presented at the American Educational Research Association Annual Meeting, Toronto, Canada. March 1978.

Scott, T.C., Meliza, L.L., Hardy, G.D., and Word, L.E. REALTRAIN validation for armor/anti-armor teams. ARI Research Report 1204. 1978.

Sells, S.B. Ecology and the Science of Psychology. Multivariate Behavioral Research. 1966, 1, 131-144.

Simon, Julian L. Basic Research Methods in Social Science. New York: Random House, 1969.

US Army Infantry School, Squad combat operations exercise (simulation), Training Circular St 7-2-172, 1975.

Wagner, H., Hibbits, N., Rosenblatt, R.D., & Schulz, R. Team Training and Evaluation Strategies: State of the Art. HumRRO Technical Report 77-1. February 1977.

Wherry, R.J. The control of bias in ratings: VII. A theory of rating. PRB Report 922. Columbus, Ohio: The Ohio State University Research Foundation, 1952.

EVALUATION OF THE MODIA PLANNING SYSTEM

INTRODUCTION

Capt John R Welsh Jr, USAF - ATC Tech. Appl. Ctr

Rand Corporation initially designed the MODIA (Method of Designing Instructional Alternatives) system as a research tool. Air Training Command (ATC) has examined the potential of the system as a computerized planning tool for use in facilitating course planning. The primary objective of the MODIA system is to provide a systematic process of relating quantitative resource requirements to course design and operation. MODIA was designed to enable planners to consider different sequences of course objectives, alternative sequences of subject matter, varying teaching methods and types of instruction, and different mixes of students, equipment, and facilities. Moreover, MODIA simulates the way in which students progress through alternative course designs. The MODIA planning system has four components: the description of options for course design, the User Interface (UI), the Resource Utilization Model (RUM), and the Cost Model (MODCOM). The UI is the interactive portion of MODIA, the RUM provides the simulation of course operation and the MODCOM provides course costs.[123]

The initial development of MODIA was completed in October of 1973. ISD teams from Keesler and Lowry performed a critical design review at that time. Rand Corporation made several revisions based on the design review, and the Phase I service test of MODIA was conducted at Keesler AFB from March 1976 to June 1976. The results of the service test were reported in ATC Project 76-1 (30 July 1976).[4] The results generally indicated that MODIA, "has the potential to be an effective planning tool whose use could lead to more cost-effective technical training courses." Several important questions, however, could not be addressed in the Phase I evaluation. This study provides data relevant to those unanswered questions.

During Phase I, Rand personnel reached the conclusion that the MODIA system was too complicated to be used effectively by the planners themselves, and as a result, a team of individuals was trained in the operation of MODIA. This group, subsequently called the interface team, operated MODIA while the course planners provided the planning data needed to design courses. The concept of the interface team carried over in this evaluation.

The physical arrangements at Keesler during Phase II were very similar to those arrangements which existed during Phase I. The special features of these arrangements included: (1) a room in which the interface team operated a remote terminal; (2) a Class A telephone line used with an acoustical coupler; (3) a MODEM (Bell 103A Data Set) in the computer facilities connected on a dedicated line to Biloxi, Mississippi telephone exchange; (4) one of the primary provisions of both Phase I and Phase II tests was that they be conducted on a "non-interference" basis. The hours of operation for the User Interface Program were to be from 0530 - 0700 hours daily, 1100 - 1200 hours - 2 days a week, and occasionally as other use dictated. While this schedule was the best that could be devised under the conditions of the service test, it placed severe restrictions on the response time of MODIA planning of alternatives and hampered evaluation of the MODIA system in that not as many course alternatives could be generated as were desirable.

The shakedown and debugging of the MODIA system on the Keesler H-6060 took place in October 1977. The actual service test for Phase II began on 14 Nov 77 and ended 17 Feb 1978.

The Phase II evaluation capitalized on the experience gained in Phase I, while expanding the scope of the evaluation of MODIA by addressing new questions about its use: in planning specific types of courses (family group courses); in controlling the system by technical school management; in assessing the value of the system to planners and managers; and in determining the data automation requirements of the system now and in the 1980s.

The objectives of this service test were to:

a. Provide sufficient test data to support the development of a Data Automation Requirement (DAR) should MODIA be adopted.

b. Determine MODIA's usefulness as a planning tool. This objective had several aspects. Specifically:

(1) Explore MODIA's usefulness in planning type 3, family group courses with shared resources.

(2) Assess the utility of the system to the technical training school management.

(3) Determine MODIA's usefulness given currently existing resource constraints, current computer support capability, and training policy.

c. Determine MODIA's usefulness as a problem-solving tool.

d. Determine the organizational configuration and operational procedures which may be used in applying MODIA at Keesler.

e. Determine resources required to implement MODIA at Keesler in the immediate future.

f. Determine what changes to MODIA are needed to improve its effectiveness.

g. Examine the potential for using the cost model (MODCOM) as a stand-alone system.

6. Develop a training program for the use of MODIA - including development of a "User Interface Team Guide".

METHOD

ISD Team Make-up. Initially, it was planned that each Instructional Systems Development (ISD) team would be composed of a curriculum training specialist, a training resource specialist, and a subject matter specialist. In practice, however, the interface team member from each of the three training groups involved in this exercise worked with only one other person from the training group. This person provided the primary ISD input. Others were involved as needed in the planning of different parts of a given course (for example, the planning of 3ABR32831 involved up to as many as 5 ISD people). The reorganization of the technical training center and shortage of experienced planners dictated this deviation from the evaluation plan. It should be mentioned that the use of fewer people significantly drove down the personnel cost of planning with the MODIA system. In contrast to Phase I cost analysis which included personnel costs of many ISD team members, this service test figures personnel costs associated with only one or at most several (in the case of 3ABR32831) ISD team members' time. Each ISD member was responsible for revising the selected courses with inputs from other tech school personnel as needed.

The MODIA Interface Team. The interface team was composed of a GS-11, a Master Sergeant, and a Technical Sergeant. The training of the interface team was accomplished at the Rand Corporation, Santa Monica, California, during the period 12 Sep 77 to 25 Sep 77. This group of individuals served as the interface between the ISD planner and the MODIA system.

Course Selection. Courses selected for MODIA service test during Phase II were:

Training Group	Course Number	Title
3380 TTG	3ABR32831	Avionics Nav System Specialist
3390 TTG	3ABR27630	A C&W Systems Operator (Manual)
3390 TTG	3ABR27630-001	A C&W Systems Operator (SAGE)
3390 TTG	3ABR27630-002	A C&W Systems Operator (4C7L)
3410 TTG	3ABR30434-1	Ground Radio Equipment Repairman (Titan)
3410 TTG	3ABR30434-5	Ground Radio Equipment Repairman (Minuteman)
3410 TTG	3ABR30434	Ground Radio Equipment Repairman

Course Selection was based on the following conditions:

(1) There were two sets of family group courses (27630 and the 30434 courses) that had to be revised.

(2) Low, medium and high student loads were represented.

(3) Different instructional approaches were represented.

(4) All courses were planned through all 5 steps of the ISD process.

(5) One of the courses (32831) was of long duration and used a great many resources.

The assumption underlying this course selection was that these courses represented the best planning possible by conventional means. If MODIA could be effectively used in the ISD process, then both planners and managers could improve course designs by allowing for more cost effective planning.

MODIA Service Test Costs -- Data Collection.

a. It is important to make a distinction here between Phase I and Phase II. Phase I results showed that MODIA could be used by training branch and group level personnel to decrease course costs through better design if they could use any design they chose, regardless of training policy or personnel management consideration. Phase II attempted to see how well they could use MODIA to manage costs in the present training environment, and within real-world constraints. One of the goals of the Phase II effort was to examine the cost of the MODIA system in the light of such restraints. In order to accomplish this goal, all elements of the system cost were collected as outlined in the evaluation plan.

b. Mnpower, facilities, equipment, and computer costs were collected by KTTC/TTGH. Total equipment and manpower cost breakouts, by course, are provided at Appendix 1. These costs will be discussed in the Results Section.

c. The primary cost of MODIA was in the manpower and computer time required to support the system. This data was gathered through work logs/time sheets and a terminal log kept by interface team members and ISD team members throughout the course of the service test. The work logs/time sheets were filled out on a weekly basis to insure current and reasonable estimates of time spent on various portions of the MODIA service test. Course cost data and cost information for use as inputs into MODCOM were provided by the Comptroller and from Keesler Production Analysis.

d. Requirements for computer resources (CPU time and time-sharing storage requirements) for the various portions of the User Interface are provided at Appendix 2. However, several recommended changes to the MODIA programs are probably extensive enough to significantly alter the operating characteristics of the system.

e. Specific changes to the MODIA system were provided in written comments by ISD personnel and all interface team members, as well as training managers.

f. Down-time, equipment malfunctions and waiting time were not counted as a direct cost of MODIA planning, since it was assumed that such costs would be minimal with a fully operational MODIA system.

Questionnaires. Data contained in responses to the questions in all questionnaires were summarized and consolidated to provide opinion information on MODIA's usefulness as a planning tool and as a problem-solving tool. Additionally, the information provided from questionnaire comments provided a basis for recommended changes to the MODIA system.

RESULTS

General. In response to one of the recommendations in ATC PR 76-1, Evaluation of the MODIA system, one of the primary purposes of this evaluation was to determine the utility of the MODIA system to technical training management. In fact, the Phase I report went so far as to say that when procedural questions and organizational configuration questions were resolved, "It appears that SAAS management will be able to show that MODIA can improve resource management in a technical training environment (para 18g, p. 33, ATC PR 76-1)." The results of the Phase II evaluation dictate a different conclusion. The next section will first discuss the implementation and operating costs of the service test at Keesler, and then report the results as they relate to the objectives previously outlined.

MODIA Phase II Service Test Costs. For the Phase II service test implementation, it cost Keesler Technical Training Center $11,074. For the operation of the system during the service test, it cost the technical training center $44,297. (See Appendix 1 for complete cost breakdown by course). These figures are considerably different from those obtained during the Phase I evaluation. For example, the Phase I report placed the implementation cost at around $36,000. The approximate $25,000 difference between that service test and this one can be explained by taking into account several important differences between the two service tests. These differences include a drastic reduction in the number of personnel and manhours involved in the shakedown and set-up phase of the service test. Additionally, fewer manhours were needed to supply interface team members with planning factors - i.e., there were fewer people involved with day-to-day operation of the system. More about these implementation and operating cost differences will be mentioned in the discussion section. Overall, the total service test cost was less than anticipated, despite the fact that computer costs were substantially greater than those costs obtained in Phase I.

MODIA System Operating Characteristics and Limitations. As mentioned in the Introduction Section, the size of the UI portion of the MODIA system was so large that it caused some initial interference with the training mission -- and resulted in restricted operating hours for the MODIA service test. The primary result of the service test experience that pertains to the operating characteristics is that the UI (and especially the "C" phase) is much too large. While it is conceivable the UI could be made smaller and more efficient, serious consideration would have to be given to size trade-offs involved with the RUM program. This trade-off is discussed in the Results Section.

In general, the operating problems experienced during the service test can be traced to the fact that MODIA programs were written for an IBM 370/158, and were somewhat incompatible with the Honeywell 6060 system. Another important factor in the incompatibility is that the IBM system is a virtual storage system, while the Honeywell system is a segmented storage system which uses program overlays. Aside from some basic incompatibility between the MODIA software and the Honeywell system, there were other problems encountered in using the UI.

Several specific findings regarding the operation and use of the MODIA programs were garnered from interface team members' responses in their questionnaire. Most team members found it easy to assign learning objectives (course content) on the UI with some notable exceptions. First, MODIA was unable to handle assignment of course content under the family grouping concept. The 250 learning event limit was much too restrictive for planners to use MODIA for simulating courses with shared resources. Two of the basic courses planned in this exercise did have 221 and 241 learning events each. Basic, single courses could easily fit within the limitations. However, courses planned under family grouping required up to three times the 250 learning event limit. Second, garbling prevented the assignment of lesson objectives which had certain letter/number combinations. For example, in the 3ABR30434 course, the interface team member entered sub10309 for a learning objective, but the computer read 1DANTCU. This garbling was a factor throughout the service test. Nevertheless, from responses to the questionnaire, it appeared that the interface team was sufficiently trained to be able to handle most problems that arose in assigning course content.

The assignment of resources to specific learning events, however, was a different matter. In all cases, interface team members found it difficult, and in some cases, extremely difficult to allocate resources to learning events in the way they wanted to make the assignments. The main problem encountered was in the extremely limited number of different resources allowed on MODIA (only 30). All interface team members had to "package" resources in order to make resource assignments using the UI. In some cases, a considerable number of resources had to be lumped together or packaged in order to make the resource assignment. In the case of planning both the 30434 and the 32831 baseline courses, very few of the needed resources could be assigned in such a way as to depict realistic use of resources.

The limited capability of MODIA to handle the required resource assignment was commented on frequently. In fact, the interface team members felt that the 30 resource limitation hampered realistic simulation of course operation since the 32831 basic course required the assignment of 80 different resources in the 30434 baseline course, over 100 different resource assignments were needed.

Another major constraint of the UI program was its inability to handle certain student arrival options. For example, staggered entries with variable numbers of entering students could not be simulated. Moreover, all interface team members voiced the need for a system which could realistically depitct shift operations. While the MODIA system can be manipulated to allow simulation of courses with a shift operation, the resultant product had severe limitations. Specifically, the 250 learning event limit for the UI was much too restrictive for depicting courses with shift operations. For example, Course 3ABR32831 would require approximately 400 learning events to simulate shift operation. Resources and learning events for this course would have had to be condensed or packaged even more to simulate a two-shift operation. Moreover, course managers felt they had a better handle on managing shift operations with current techniques.

All interface team members felt confident in using the UI and all found the User's Guide (provided by Rand) very helpful in working the system. But again, there was considerable difficulty in working around garbling problems. As in Phase I, the numbers 143 and 168 were read as 145 and 170. This particular garbling caused a problem every time one had to enter learning event numbers 143 or 168. The problem was surmounted by labeling these learning events as "sick-call" and assigning zero time to the learning event.

In general, the interface team felt confident with the simulations and expressed the need for a system like MODIA, but all members also remarked that the system, in its present form, had too many limitations. All interface members rated the output of the RUM as fairly easy to interpret and use, and of value in the planning process.

The specific changes recommended for MODIA are dealt with in detail later. Suffice it to say that in the experience of the Phase II service test, the UI system was too large and inefficient to be used on the Keesler H-6060 computer now, or in the future. Based on limitations and problems experienced in this portion of the service test, and on the results reported in the next section, MODIA should not be adopted for us "as is". The operating limitations experienced during Phase II were very similar to those experienced in Phase I.

MODIA Usefulness as a Planning Tool

General. The basic thrust of the Phase II evaluation effort was to determine MODIA's usefulness as a management planning and problem-solving tool. The assessment of MODIA's utility must, of necessity, be subjective and depend on the opinions and judgments of those in the training center management hierarchy who would use a system such as MODIA. The strategy of this evaluation effort was to present course managers with MODIA products (the RUM simulation and Cost Model course costs for each alternative design) and see if they could use either the simulation information or cost model information to arrive at more cost effective course designs -- while staying within the limits of command and center level policy directives, manpower limitations and resource constraints.

The basic question involves "how" management should use the MODIA system. Therefore, management responses and results of alternative course costs will be discussed with respect to obtained results in each of the technical training groups, respectively. To help the reader keep this part of the evaluation conceptually straight, there are two basic aspects of MODIA information that could be of use. The questions that address these aspects are: (1) How useful was the cost model information on alternative course designs? and, (2) How useful was the simulation information? The first question is answered by results discussed here, while results pertaining to the second question are discussed later. In the discussion to follow, each of five basic courses in the technical training groups was simulated and cost for the courses calculated using the cost model. The "baseline" course cost figures and operating parameters were designed to reflect the way the course actually operated during 1977. All baseline course costs were figures on the 6-hour training day and were compared with total course cost figures derived using the ATC Comptroller's figures on costs/graduate in each of the training courses multiplied by the annual graduates. The results show that the cost model figure for total course costs agrees closely with a total course cost using the comptroller's cost factors -- a result that agrees with Phase I findings on cost model accuracy.

Because of the time limits of the service test, cost model information was obtained for three basic courses and alternatives for each of the three courses.

3380 Technical Training Group - Results of Alternative Costing.

Table 1 presents the cost of the baseline 3ABR328131 course in comparison with costs of various alternatives generated by the interface team members, ISD people, and training managers.

While the cost model was designed to be used either in conjunction with the simulation portion or by itself, ideally, planners would cost out alternative courses they had simulated to determine the most cost effective option. In the present case, five out of eight alternatives were simulated.

Referring to Table 1, alternatives 7 and 8 were less expensive than the other alternatives. These two courses represented slightly increased

TABLE 1

COST MODEL BASELINE COURSE COST COMPARISONS WITH ALTERNATIVE COURSE COSTS (3ABR32831)[a]

Alternative No.	Baseline[c]	1[d]	2	3	4	5	6	7	8
Data									
1. Annual Entry	412	486	486	730	1128	1128	389	389	389
2. Annual Graduates	269	318	477	477	737	737	255	266	282
3. Annual Failures	7	8	13	13	19	19	6	6	6
4. Avg Course Hours	527	515	527	527	527	527	527	616	537
5. Number of Instructors	32	34	59	52	73	63	38	38	38
6. Total Course Cost (1977 dollars)[b]	1882.9	2137.4	3383.3	3279.4	5000.2	4857.0	1912.0	1802.6	1708.0
7. ATC Course Cost/Graduate	7.23205								
8. ATC Total Course Cost ((7)X(2))	1945.4								
9. Difference (Between 8 and 6)	63.5 (3.2%)								

a. In thousands of dollars
b. Adjusted by a factor of 1.1598 X MODCOM value X 1.057 = 1977$
(Factors provided by Hq ATC/Management Analysis)
c. 6-hr day (Baseline)
d. 8-hr day

course hours over the baseline course (the course as it actually operated on a 6-hr day and for 527 course hours). The modest reductions of $80.3K (4.3%) for alternative 7 and $173.0K (9.21%) for alternative 8 were obtained by entering students every 36 hours instead of every 40 hours as in the baseline, and by decreasing the actual numbers of students entering the course. Additionally, students were washed out on an average of 217 hours in alternative 8, as opposed to an average of 240 hours to washout in the baseline. The reduction in alternative 7 was due largely to reduced student pay, reduced student PCS costs and instructor PCS costs.

All training managers felt the cost information provided by the cost model was of very little use to them. All sampled managers commented that the cost model information could be used by Hq ATC level people involved in making decisions about policy impact on course costs. The managers in this group indicated that the cost savings shown by using the cost model related to costs not managed by center level managers (student instructor pay and PCS costs). These obtained cost savings were in areas most directly controlled by Hq ATC management actions.

It is significant that the two money saving alternatives (both 7 and 8) were generated independently of any simulation -- i.e., neither of those two alternatives were put on the RUM. This fact demonstrates that the cost model may be used as a stand-alone system, but the overriding question as to "who" should use it is addressed later in this report. Because of the recommendations generated by training managers in this group as well as others, the cost model data was given to Hq ATC Comptroller personnel in the management analysis section for further study and comment

3390 TTG - Results of Alternative Costing.

Table 2 presents the cost of the alternatives run for the 3ABR27630-000 course. In this case, the baseline course cost generated by the cost model was within 7% of the ATC Comptroller's figures for the cost of the basic course in 1977. The second alternative simply represents the course cost based on an eight-hour training day as opposed to a 6-hour training day. The resultant savings are trivial ($4.2K). The remaining alternatives represent various ways of figuring course length based on how policy dictated reductions are calculated. While the cost model may give the cheapest alternative (Alternative 2 in this case), planners still needed to exercise judgment. The vast bulk of the savings generated by this alternative was in pay and allowances of students, instructors, and base permanent party (support) personnel. These costs, while important, are not meaningfully controlled by managers at the training group level. Additionally, the cheapest alternative is not always the best. In the base of 3ABR27630-000, the Complementary Technical Training (CTT) is vitally important to the course of training. Alternative 4 presented the planners with the best course length and number of graduates from the standpoint of meeting training standards. It is important to note that the planners and managers

TABLE 2

COST MODEL BASELINE COURSE COST COMPARISONS WITH ALTERNATIVE COURSE COSTS (3ABR27630-000)[a]

Alternative	Baseline[c]	1[d]	2[e]	3[f]	4[g]
Data					
1. Annual Entry	550	550	550	550	550
2. Annual Graduates	471	471	490	471	481
3. Annual Failures	11	11	11	11	11
4. Avg Course Hours	215	285	215	274	242
5. Number of Instructors	25	25	25	25	25
6. Total Course Cost[b] (1977 dollars)	1717.1	1713.9	1567.4	1687.1	1622.7
7. ATC Course Cost/Grad	3.92178				
8. ATC Total Course Cost	1847.2				
9. Difference (between 8 and 6)	-130.1 (7%)				

a & b - Same as other tables
c. 6-hr day
d. 8-hr day
e. 8-hr day (not adding CTT)
f. 8-hr day course + (CTT - 15%)
g. 8-hr day (Course + CTT) - 15%

TABLE 3

COST MODEL BASELINE COURSE COSTS COMPARISONS WITH ALTERNATIVE COURSE COSTS (3ABR30434) [a]

Alternative #	Baseline	1 [c]	2 [d]
Data			
1. Annual Entry	591	591	591
2. Annual Graduates	402	424	435
3. Annual Failures	25	26	26
4. Avg Course Hours	460	520	460
5. Number of Instructors	73	73	73
6. Total Course Cost (1977 dollars) [b]	3280.2	3105.1	2986.9
7. ATC Course Cost/Graduate	8.46088		
8. ATC Total Course Cost ((7) X (2))	3401.2		
9. Difference (Between 8 and 6)	- 121.00 (-3.5%)		

a. In thousands of dollars

b. Adjusted by a factor of 1.1598 X MODCOM value X 1.057 - 1977 dollars (Provided by Hq ATC/Management Analysis)

c. 8-hr day - same course length

d. (Course + CTT) - 15%

already knew that the mix represented in alternative 4 would be the best option for them to plan under the new policies. It is also important to emphasize that the way in which these policy decisions would be implemented was determined independently of cost model information. The cost model data confirmed what managers already knew about the effect of recent policy decisions on course costs.

Both managers in this group rated the cost model information of no use at all to them. They indicated that the information could be of use to Hq ATC Training Managers and others in evaluating the cost of current course training and in evaluating the cost of alternative course designs.

3410 TTG - Results of Alternative Costing. Table 3 contains the results of alternative costing for the 3ABR30434 course. Only two alternatives were run through the cost model in this course because time ran out for the service test. A problem with the amount of time necessary to gather data for input into MODCOM (cost model) bears examination at this point. This problem occurred with gathering data to input into all three baseline courses, but is discussed here for convenience. All interface team members spent a great deal of time gathering cost information and putting it in a form usable by the cost model. Specifically, for the 3ABR30434 course, 30% of the total time spent by the interface team members for all phases of the service test was in gathering cost data for MODCOM (for interface team members on 3ABR32831 - 50% and 3ABR27630 - 50%). This amount of time is grossly disproportional when one considers the inability of managers to use the final cost information. In any event, the interface team felt that entirely too much time was spent on this portion of the service test. The mechanics of inputting the information and obtaining final products was, on the other hand, extremely easy and presented no problems whatsoever in terms of usage. Once the baseline cost information was obtained, it was very easy to generate costs for alternative course designs.

The two alternatives presented for the 30434 course show a roughly 5% saving (alternative 1) for the course planned on a straight 8-hour training day with no CTT added; and a roughly 9% saving when planning a course length with CTT added and reducing the resultant course hours by 15%. As with the other training groups, this information was not enightening to training managers. One of the managers found the cost model information very useful in improving course manning structure and student use of equipment. Both managers felt that costs were generally determined by ATC policy, rather than managed by training group-level personnel.

MODIA's Usefulness as a Problem-Solving Tool -- The Resource Utilization Model Simulation. Based on the responses to questionnaires (see Appendix) given to the training managers in the three training groups involved in the Phase II service test and on interviews with group and center-level management, the RUM does not appear to be a useful planning tool that can be used by these managers to more effectively manage training resources.

General. As mentioned in the Methodology section, all managers were presented with completed course simulations and cost model data for all alternatives -- all managers had the products carefully explained to them (8/9 of the managers responded that they had the products explained well enough to them that they understood all the products from the RUM and the cost model). All managers were facing real course management problems at the time of the evaluation. For example, all had to revise courses from a six to an eight-hour training day, all had instructor shortages, and all had students awaiting training. The RUM simulation was unable to provide the training managers with unique information on course operation. From their responses to questions 1, 2, 9, 11, 12, 13 14, 15, 16 and 17, it was apparent that the MODIA simulation was not telling the managers something they did not already know about specific problems in the operation of the courses studied in this service test. It appeared from responses given in the questionnaire that the RUM simulation was generally of very little use to the group level managers. From responses generated during debriefing and from comments on the questionnaires, the problems concerning the managers were foreseen without the aid of the simulation. The more pressing problems, such as those concerned with shift operation, could NOT be realistically simulated on MODIA. Seven managers said they had little confidence in the RUM simulation, and all nine managers felt they had foreseen the problems in course operation just as well or better than the simulation.

At this point it is of interest to note that enthusiasm for MODIA ran high during the service test because of perceived potential of the system for helping managers solve some of the problems that were facing them at the time. However, the managers expressed frustration with the MODIA system when they could not use it to help them manage those problems. For example, one branch chief would have liked a system which would allow him to strategically pull instructors to support Air Force exercises and still optimally operate the course.

At the conclusion of the service test, several managers expressed the need for a computerized system which would help them with scheduling problems, and/or a system which would optimize the use of certain resources. It was explained to the managers that MODIA was neither a scheduling nor an optimizing tool. One has to use the RUM simulation to test the feasibility of the given design the planner brings to the system. Other systems, such as the Advanced Instructional System (AIS), could be used to resolve the scheduling and optimizing problems which seem to represent the more important management problems facing course managers at the center and group level.

1411

The results of the managers' opinions about the simulation differed slightly from the opinions of the ISD participants and from the judgments of the interface team members. Two-thirds of the ISD people felt the RUM simulation would be of little use in course planning, while two-thirds of the interface team members felt it would be useful. Both ISD respondents and interface team members were confident in the simulations of their respective courses. Of the six respondents (3 ISD people and 3 interface team members), three said they would seldom use MODIA were it to become a fully operational system, and three said they would use the system often. This result compares to the opinions of the training managers, where five of nine said they would at least use the system "sometime" if it were fully operational.

In summary, the opinions of those involved with the Phase II service test found the simulation lacking in certain respects. In general, there were mixed feelings about the usefulness of the RUM simulation. The managers felt that the simulation was of little value to them, but the ISD and interface team members were of the mixed opinion that perhaps there was some value to be had in the way MODIA simulated course operation. All individuals sampled with the questionnaire felt that the simulation of resource use was less than totally realistic and those most closely involved with MODIA expressed serious reservations about the restrictive limits on the number of resources that could be planned using MODIA. A summary of the training managers' responses in each training group is presented below.

The 3380 TTG - 3ABR32831. The interface and ISD participants for this training group had some difficulty packaging resources for this long duration and high-flow course. For example, there were 80 important resources that could not be broken out as desired in the basic course simulation. In fact, the course structure as it existed in actual operation could not be accurately depicted. 3ABR32831 could not be depicted as progressing from group lock-step, to a self-paced portion, then back to a group lock-step again. The specific problem facing the simulation of 3ABR32831 was that the students had to be returned to that portion of the self-paced block from which they were taken in order to complete the last group-lock-step block. The configuration that was simulated had a self-paced portion at the end of the two lock-step portions with students arriving in random intervals. The training managers who examined the simulation felt that such a simulation was of very little use or no use at all to them -- 5/5 responses were in this category; 4/5 of the managers in this training group felt the simulation and cost model information were of little value; and 3/5 had little confidence in the final simulation.

The 3390 TTG - 3ABR27630 Course. The two training managers in the 3390 Technical Training Group found MODIA to be more useful than did the managers in the other two training groups. Both felt MODIA would be useful in helping them manage course revisions better, both thought the system had value to them as managers, and both were very confident in the results. The difference in the response of this group and the others can be attributed to the fact that MODIA simulation helped the managers spot a queuing problem that existed in the 27630 course operation. While the managers knew a problem of some kind existed, it seems MODIA highlighted a possible solution which was subsequently put into operation -- the queuing problem was solved.

While MODIA generally provided favorable results in this training group, several comments by the managers are important in assessing MODIA's usefulness. First MODIA could not adequately simulate the group-paced operation of the course. MODIA, however, can be manipulated to handle group-paced instruction, but in the present case the options available for simulating group-paced instruction were not acceptable to training managers. The managers were not satisfied with the way the resultant course "looked" in the simulation. Second, MODIA didn't allow the managers to more effectively manage resources. As the branch chief remarked, "In its present form, the only useful purpose it serves is to highlight the facility costs in one single document." The most pressing problems facing these managers were instructor manning shortages. They felt it would be futile to exercise a system that merely highlighted the manning problems they were aware of already.

The 3410 TTG - 3ABR30434. The most predominant remarks made by managers in this group dealt with the limitations of the MODIA system. Both managers felt that MODIA in its present form would be of little value and they had little confidence in the simulation. They both said that they would use MODIA often, if it were substantially changed.

Organizational Configuration, Operational Procedures, and Resources Required to Implement MODIA. As stated in the MODIA Evaluation Plan for the Phase II service test, the determination of how MODIA should be used largely depended on how well the planners and managers felt they could use the simulation and cost information. The results of the two preceding sections indicate that the RUM simulation of course operation was of too limited a scope to be of any value in the planning and management of course operation at the center, group or branch levels.

All personnel involved in the service test were queried as to how and where MODIA should be used if it were adopted. There was a wide range in the recommendations as to who should use MODIA. Some managers felt that only branch level planners should use the system, while most others recommended use by everyone involved in the planning process, from branch level to Hq ATC course managers and manpower personnel. Many managers stated they would not recommend the system as it exists now, but stated that they could use a MODIA-like system. A particular surprise was the suggestion by many that Hq ATC level personnel could use cost model information.

This recommendation was surprising in light of the fact that at the beginning of the service test, managers and other training center personnel expressed a fear that MODIA would be used by headquarters to impose unrealistic course policy changes on them. That training managers thought enough of the cost model to recommend its use by Hq ATC planners speaks well for the cost model. Again, though, almost all the personnel questioned did emphasize that the simulation could not be used by them unless it was considerably changed.

The comments and responses about the best organizational configuration were clear in the recommendations that one centrally located interface team could handle the planning of all Type 3 courses at Keesler. By far the most frequent recommendation was that a well-trained interface team composed of only 3 members could handle all the necessary planning.

In addressing the questions of resources required to implement MODIA, the results of this service test have several clear implications. As far as the manpower required to operate the system, the results of this evaluation indicate that very little manpower increases would be needed to operate the system effectively. This result is consistent with the most prevalent recommendation in this service test -- that MODIA be operated by a centrally located team of about 3 individuals. The unexpectedly low cost of this service test was achieved for a variety of reasons, dealt with fully in the discussion section, but in general the results indicate much fewer manhours involved in operating the system than may have been estimated based on Phase I results.

In contrast to the small manpower increases that would be required to implement MODIA, the results of the service test indicate considerable expenditures in computer resources would be required to implement the MODIA system.

The severe interference with training caused by operating the UI and the resultant restruction in operating hours for the service test indicate that MODIA as it currently is written could NOT be used for eight hours a day without causing unacceptable impairment of other training being conducted on the H-6060. There appears to be little possibility of using MODIA on the B-3500 system, since Hq ATC/ACD has gone on record as stating the B-3500 system is currently saturated. The computer personnel at Keesler felt MODIA could not under any circumstances be used in its present form, since existing computer resources and current training priorities leave little room for a system as large as MODIA. The unacceptability of MODIA "as it currently exists" is a consistent theme that runs through the comments of all those involved with the service test. The recommendations advanced for making MODIA more acceptable and usable are discussed in the next section.

Necessary Changes to Improve MODIA Effectiveness. One of the standout results of this evaluation was that MODIA would have to be dramatically changed if it were to be an effective planning tool. Far and away the most prevalent recommendation for change in the RUM was that the limit on the number of training resources be considerably increased. The current limit of 30 resources is just not adequate. All courses planned in this service test, as well as one of the courses in the Phase I evaluation, had difficulty working around this limitation. The magnitude of the problem created by a limit of 30 resources is highlighted when one considers that the average number of resources used in most courses is considerably larger than 30 (and can go as high as 1100 resources in one particular course).

In addition to increasing the resource limitations on the UI and RUM portions of MODIA, it is necessary to decrease the overall size of the UI program, especially the "C" phase of the UI. This phase requires 70K bytes of storage in a time-sharing system with 110K bytes available for users. The large portion of the time-sharing system required by this phase causes unacceptable interference with other users of the time-sharing system. This particular recommendation for reduction in the size of the UI is a result which was also obtained in the Phase I service test.

At this point it must be mentioned that while the UI could be rewritten to be more efficient and still handle the recommended size increases discussed below, the resultant size increases in the operation of the RUM would probably prove unacceptable. Since the limits that apply to the UI directly affect the amount of storage required by the RUM, increases in the limits allowed on the UI would greatly increase the amount of computer time and core storage demanded by the RUM.

In relation to the problem of the overall size of the UI program (especially the 'R' and 'C' phases), the interface team recommended that provision be made in the program to enter a given phase at particular points during the phase. As the programs are currently written, the user must enter a phase at the beginning if he/she is to make a change and the user must go through the entire phase and reenter all subsequent phases. As a consequence, a considerable amount of time can be expended for relatively minor changes in the resource assignments or capacities. Aside from having to reenter all subsequent information and the amount of time and effort involved, having to tie up the computer for relatively minor changes impairs the cost effectiveness of the UI.

While shift operations can be planned on MODIA as it currently exists, the programs should be rewritten so as to allow more direct simulation of the shift operation. Current options on MODIA necessitate manipulation of the MODIA system in such a way as to make simulation of shift operations unrealistic and unacceptable to training managers.

Apparent form the garbling of certain numbers and letter combinations and the large size of MODIA programs was a certain amount of incompatibility between the MODIA software and the Honeywell system. MODIA programs currently cannot be rewritten to alleviate this basic incompatibility, but should MODIA eventually be adopted, the garbling problem would have to be resolved.

DISCUSSION

General. The Phase II evaluation effort differed from Phase I in that MODIA planning was attempted in an operating environment with planners and managers judging and considering MODIA simulation and cost information in the light of current policy guidelines and resource constraints. The results of their judgments and the operating experiences of this service test indicate that the simulation has very limited value for the management of technical training courses at the branch, group, and center level. This result is at apparent odds with Phase I findings which seem to indicate MODIA had potential for helping management design more cost-effective courses.

In the Phase I effort, however, the primary concern was determining whether the simulations of course operations were "valid", i.e., could they realistically simulate the way courses actually operated; and whether the cost of MODIA could be offset by more cost-effective designs -- as compared with conventionally designed courses. Phase II mainly tried to determine how course managers might use the system. The results clearly indicate that the simulation and cost information was of little use to managers at this level.

No attempt was made to compare MODIA planning with conventional planning. Such comparisons are a little like comparing apples and oranges. Were MODIA to be radically changed, the interface team and 2 of 3 ISD planners felt the system could be an important aid in organizing and clarifying the planning process. In short, MODIA could have been an important addition to the planning process, but limitations in the system as it currently exists precluded managers from seeing it as a positive, useful tool. Specific highlights of the results will be discussed in relation to Phase I results and in relation to MODIA's potential for implementation now and in the future.

MODIA Costs. One of the most striking differences between Phase I and Phase II service tests was the costs of implementing and operating the MODIA system. The main difference in the cost of the two service tests was in the reduced manpower required to operate MODIA. This reduction reflects the high level of competence achieved by the interface team. The experiences of Phase I seemed to have paid some dividends for the Phase II effort. Particular mention must be made of the quality of training of the interface team. The interface team members, as well as the project officer, were very adept at working around problems and seemed to know a great deal more about how to use the system than may have been the case in Phase I. All four individuals expressed complete satisfaction with the thoroughness of the training received at Rand Corporation in Santa Monica. The monies expended for this training undoubtedly paid big dividends both in the reduced cost of operating the system and in the quality of the generated products.

Just how much this experience of the service test would affect the per hour cost of MODIA planning figures in the Phase I effort is difficult to guess exactly since this service test did not address the cost of MODIA planning vis-a-vis conventional planning. It is safe to say MODIA planning would not be as expensive as indicated in the Phase I test. The carefully kept work logs and the comments by virtually all participants (Interface Team, ISD, and Training Managers) indicated MODIA planning would not unduly complicate the planning process. They felt the simulation would be a useful aid in course planning were its limitations corrected to allow more realistic representation of resource use and more realistic simulation of course shift operations. The limitation of the UI and the RUM products appears to be the main factor mitigating against MODIA effectiveness.

MODIA System Limitations. A look at the recommended changes to MODIA provides the reader with a base from which to judge the limitations of the system. The standout limitation is the restricted number of resources that can be assigned to learning events. This limitation probably reduced the acceptability of the simulation to the majority of the training managers. Again, this finding is unenlightening to some extent since Phase I findings indicated that the limitations degraded its acceptability to course planners. Phase II results showed the limitations degraded the system acceptability to managers as well.

In addition to the resource limits, the interface team reiterated the desire to have a user interface that would be more adaptable. They indicated a need for a system which could be entered at more points in a given phase, and which could accept changes within a phase without having to reenter all subsequent information within a file.

Another major factor influencing acceptability of MODIA was the inability of planners to realistically simulate courses with shift operations, and courses with certain configurations. 3ABR32831 had certain portions of the course where students progressed from lock-step through self-paced instruction and then back to lock-step again. The course could only be simulated with the self-paced portion at the end of the two lock-step blocks. While this resolved the problem of simulating the course with MODIA, it made the resultant "picture" of course operation unrealistic. The problem of simulating courses with shift operations was more serious. Managers expressed a definite need for accurately simulating this type of operation, but the MODIA system was not designed to handle shift operations in a way that would be useful to course managers.

In planning 3ABR27630, the managers were chagrined by MODIA's inability to simulate group-paced instruction. The course was simulated using the lock-step option, but again, the resultant simulation was somewhat unrealistic and the training managers expressed their dissatisfaction with the resulting product.

In sum, while some realistic simulation of the three courses was achieved, managers felt that inherent limitations of the MODIA system prevented the simulation from being of any value to them in managing training courses. A majority of the managers liked the MODIA concept, but wanted the system to do more than it was designed to do.

As mentioned earlier, there are other scheduling and optimizing models such as AIS, which better handle the problems facing course managers.

A question naturally arises about the relatively small numbers of managers exposed to the MODIA simulation in this service test. From a rigorous standpoint, it would be unwise to generalize about the value of MODIA as a management tool based on the judgments of only nine course managers. These nine were sampled in this service test because they were most directly involved in the management of the courses selected for the service test and in the best position to judge the utility and accuracy of the MODIA simulation and cost model information. Additionally, the large reorganization of the former School of Applied Aerospace Sciences under the training center caused other managers who would otherwise be involved to be shifted to other organizational positions. Only those managers who could best judge the accuracy and usefulness were asked to comment on the system.

Arguing for the generalizability of the training managers' judgments is the fact that the perceived limitations of the system were, by and large, the same limitations uncovered in the Phase I service test by course planners. The fact that these limitations were also judged by the training managers as constraining the usefulness of the simulation to management, provides a reasonable clue as to the value of MODIA simulation to others. It can be argued here that if the managers most familiar with course operation could not find the simulation useful or acceptable, no one else could either.

The Cost Model. In general, the training management found the costing of course alternatives very interesting but of little value to them in management of course operation or in planning revised courses. The large majority of managers stated that the cost model information was of little practical value at their level, but went so far as to suggest the use of cost model information by Hq ATC level management. The obtained accuracy of the cost model figures and the ease with which alternative course costs could be generated argued strongly for its adoption at some level. The inputs to the cost model portion of MODIA can be obtained independently of any information provided by the simulation portion of the MODIA system. In short, from the results of this evaluation, MODCOM could be useful for Hq ATC level planners and managers. Based on the service test experience, however, certain caveats have to be issued regarding the time involved to gather and format information for input into the cost model.

The input information for use by MODCOM took a long time to gather and put into a usable form. Regardless of who uses the product, input undoubtedly would have to come from branch and group level planners, and would have to be updated regularly by the same people. This effort would naturally extend to all Type 3 courses at each of the training centers and would involve substantial changes in the way maintenance and cost data on training course resources are kept. Would the effort be worth it? In order to get a feel for the utility of the information provided by the cost model, the cost model information on each of the three basic courses and for all alternatives examined in this service test were given to planners in the Management Analysis Directorate of the Hq ATC Comptroller. The results of their study of MODCOM indicated that the Cost Model would probably not be of any use at the Hq ATC level.

Despite the opinion of group and branch level management that MODCOM information was interesting, but of little value, it could be argued that these managers should be using this information regardless of current practice. The argument may go that just because they are not used to considering costs of alternative course designs, they could or should make such considerations when they plan or revise courses. This position involves the determination of "what" or "who" drives course costs. The managers sampled in this exercise felt that training philosophy, as presented in center-level and headquarters-level policy guidelines (as well as practical considerations like the Trained Personnel Requirement) drove the bulk of the costs of training. As it turns out, current systems for managing course costs are adequate, and the cost model information would not add anything to current management of course costs.

In sum, the major conclusion of the evaluation study is that MODIA could not be of sufficient practical value to managers at the branch, group, or center level. MODIA was designed as a research tool to answer broad, "what if" types of questions and probably should not be modified to provide the more detailed level of simulation required by these planners and managers. In addition to the simulation output from MODIA, the Cost Model information proved to be of little value to the planners and managers at this organizational level for similar reasons. Specifically, these individuals did not manage those costs which represented the largest part of the variable costs associated with technical training.

It could be argued that planners and managers at higher organizational levels should routinely use the simulation and cost model information to examine the impact of broad policy decisions on course operation, but such usage has inherent limitations. Someone would still have to provide the baseline course data and keep it up on all courses of interest, and from the experience of this service test, this would be no small chore.

A more realistic use of the MODIA system probably involves using the system for what it was designed to do best -- answer broad research questions such as: (1) What are the effects of varying student ability on course design? (2) When is self-pacing best? - - with what types of courses? (3) What is the interaction between student ability and types of training and course cost? (4) How to best group student training on expensive equipment? MODIA may allow researchers to approach these and similar questions without extensive and expensive (and often equivocal) field studies.

REFERENCES

1. Carpenter-Huffman, Polly. Overview of a Tool for Planning the Use of Air Force Training Resources, Vol 1, R-1700-AF, Rand, 1977
2. Carpenter-Huffman, Polly. Options for Course Design, Vol 2, R-17001 - AF, Rand, 1977
3. Hess, R & Kantar, P., A User's Guide to the Cost Model, Vol 5, R-1704 - AF, Rand, 1977
4. ATC PR 76-1, Evaluation of the MODIA System, 30 July 1976

APPENDIX

KTTC COST INCURRED TO IMPLEMENT/OPERATE
MODIA PHASE II SERVICE TEST

CATEGORY	FACTOR	RATE	COST
I. IMPLEMENTATION COSTS*			
A. Computer Terminal Time	16.53 hr	$220.00/hr	$3737.00
B. Other Projects run on MODIA Systems	15.4 hrs	220.00/hr	3391.00
C. Communications line installation costs (installation/removal)			200.00
D. TDY Costs			
1. Interface Team Training.	--	--	3194.00
2. Lackland TDY	--	--	262.00
E. MODEM Costs	5 month	58.00/Mo	290.00
SUB TOTAL			11.074.00[1]
II. OPERATING COSTS			
A. 3ABR32831			
1. Computer terminal time	44.8 hr	220.00/hr	9856.00
2. Personnel usage			
1 E-6 (interface)	178	6.45/hr	184.45
2 E-5 (ISD)	47	5.39/hr	253.33
3 E-5 "	97	5.39/hr	522.00
4 E-4 "	22	4.65/hr	102.00
5 GS-9 "	22.5	9.43/hr	212.24
Sub Total			11,130.00[1]
B. 3ABR30434 (all three courses)			
1. Computer Terminal time	16.5 hrs	220.00/hr	3,630.00
2. Personnel Usage			
a. E-7	133.25	7.55/hr	1,006.00
b. GS-9	108.5	9.43/hr	1,023.00
Sub Total			5,659.00[1]

[1] Rounded to nearest dollar.

*Excludes cost of Rand personnel usage in setting up MODIA.

CATEGORY	FACTOR	RATE	COST
C. 3ABR27630 (All three courses)			
1. Computer terminal time	50.3 hrs	220.00/hr	$11,066.00
2. Personnel Usage			
a. GS-11 (Interface)	33.25	11.41/hr	379.38
b. GS-9 (ISD)	15.5	9.43/hr	146.17
Sub Total			$11,592.00[1]
OTHER COSTS			
Personnel Costs:			
Waiting (Computer Proglems - Access, etc.)			
E-7	7.00	7.55	$ 52.85
E-6	97.75	6.45	630.49
GS-11	45.25	11.41	516.30
SUB TOTAL			$ 1,199.00
TOTALS			
Project Officer (Keesler)			
Capt (69 days x 8 x .6)	331.00	11.01	$ 3,644.00
Total Computer Terminal Time Costs			$31,680.00
Total Personnel Costs (Manpower Time)			3,828.00
Total Implementation			3,946.00
Total Other			1,199.00
TOTAL COST			$44,297.00

STORAGE REQUIREMENTS[1] FOR VARIOUS PORTIONS OF THE USER INTERFACE

PHASE	K Bytes of Storage
I	42
S	42
P	46
T	40
R	58
C	70
RUM	48
RUM	66

---

[1] On H-6060 Time-sharing System

INSTRUCTIONS FOR THE TRAINING MANAGERS' QUESTIONNAIRE

1. As a manager of various aspects of technical training course operation, you have been asked to participate in the evaluation of the MODIA planning system. You are in a position to make a number of assessments concerning the usefulness of MODIA products to the effective management of training course operations, and to MODIA's usefulness to the training manager as a planning tool. In the following pages, you will find a number of specific questions concerning your experience with MODIA and your opinions as to its potential usefulness. The judgments required of you are approximate in nature, but please exercise thoughtful consideration for each question. Only summary statistical results of your responses combined with responses of others will be used in deciding on the utility of the MODIA system.

2. Please read each question carefully and indicate your response on the rating scale by placing a check mark in the appropriate space. When you are finished, make sure you have completed the general information sheet, and that you have put your name in the indicated place. Place the completed questionnaire in the envelope provided and return it to TTGOT. Remember, no other Technical Training Wing personnel will see your responses. The information will be analyzed and presented in summary statistical form by the Technology Applications Center. If you have additional comments, clarification and/or explanations, regarding any particular question, please make them on the back of the sheet containing the question. Please indicate the question number.

QUESTIONNAIRE FOR TRAINING MANAGERS

1. How useful was the MODIA simulation to you as a training manager, in spotting potential problems in course operation?

: 1 :	: 5 :	: 1 :	: 2 :	: :
NO USE AT ALL	VERY LITTLE USE	MODERATE USEFULNESS	CONSIDERABLY USEFUL	EXTREMELY USEFUL

Please list the potential problems that MODIA allowed you to spot.
- problems in instructor manning in lab situations
- Instructional sequencing
- Number of required classrooms
- No problems not already foreseen
- Student bottlenecks

2. Were any of the problems depicted in the MODIA simulation problems you would have foreseen without MODIA?

8 Yes   1 No

If your response was yes, please list those problems you could have foreseen without MODIA simulation and explain how you would have foreseen them?
- Manpower Utilization
- Delays in student progress
- Costs of Training
- Laboratory Utilization

Which problems would you have been unable to foresee without the MODIA simulation? (Please list)
- None
- All problems were known before MODIA simulation
- All of the problems established by MODIA programs were foreseen and attacked without MODIA

3. How realistic were the alternative course designs provided you by your ISD team member?

TOTALLY UNREALISTIC	SOMEWHAT REALISTIC	MODERATELY REALISTIC	VERY REALISTIC	VIVIDLY REALISTIC
2	3	2	2	

Please comment on aspects of the alternatives which you feel were helpful or unrealistic:

- The alternatives could only provide single shift simulation due to amount of inputs required.
- MODIA was limited because suggested alternatives could not be used because of the 250 learning event restriction.
- The alternatives that were used indicated the results that were anticipated.

4. How much time did it take for the ISD team to generate alternative course designs for you?

NO TIME AT ALL	VERY LITTLE	MODEST AMOUNT	CONSIDERABLE	AN EXTREME AMOUNT OF TIME
	1	4	4	

5. How many of the course design changes you recommended were the ISD interface team members able to incorporate into the alternative course designs?

NONE	VERY FEW	SOME	MOST	ALL
2	3	1	2	1

Do you have any additional comments on the alternatives the ISD teams designed on MODIA for you?

6. Could you understand the output of the MODIA simulation?

8 - Yes
1 - Not at all

1428

Comments?

7. Did the ISD team member explain the output to you?

8 Yes   1 No

8. Did you feel that you could understand the simulated course operation after it was explained to you?

8 Yes   1 No

Comments?

9. In your opinion, would the MODIA simulation of course operation and course cost enable you to better manage course problems and resources?

: 1 :	: 3 :	: :	: 3 :	: 2 :
VERY MUCH SO	SOMEWHAT BETTER		VERY LITTLE	NOT AT ALL

Comments?

- The simulation would enable us to find bottlenecks and queuing problems before they occurred. It would be extremely valuable as a course planning tool if we were able to program the inputs for a two shift course.
- Program needs expanding to allow other management factors to be considered: i.e., class schedules, washback related problems for rescheduling, etc.

10. Are there any changes you would like to see made in either the course simulation or cost information that would make the MODIA system output more useful to you as a course manager? If there are any changes, please list them and explain?

- Computer time needs to be increased.
- Limits on the number of inputs requires increasing
- Output needs to be reorganized by higher Hq as valid tool for increasing or decreasing manning and/or facilities.
- Cost data was very difficult gathering and validating.
- Increase the 250 training event limitation
- Increase type of learning events, teaching formats and teaching agents.

11. How often do you feel you would use MODIA were it to be adopted as a fully operational system?

: :	: 4 :	: 3 :	: 2 :	: :
NEVER	SELDOM	SOMETIME	OFTEN	CONSTANTLY

12. Overall, how valuable would MODIA be to you in planning a course revision?

: :	: 6 :	: 1 :	: 2 :	: :
NO VALUE	LITTLE VALUE	MODERATE VALUE	VALUABLE	EXTREMELY VALUABLE

Additional comments?

13. How confident are you in the results of the simulation of course operation?

: :	: 5 :	: 2 :	: 2 :	: :
NOT CONFIDENT AT ALL	VERY LITTLE CONFIDENCE	MODERATELY CONFIDENT	VERY CONFIDENT	EXTREMELY CONFIDENT

14. How confident are you in the cost figures shown to you on the course costs (including the alternative course designs)?

: 2 :	: 4 :	: 1 :	: 2 :	: :
NOT CONFIDENT AT ALL	VERY LITTLE CONFIDENCE	MODERATELY CONFIDENT	VERY CONFIDENT	EXTREMELY CONFIDENT

15. How useful a planning tool would the course simulation be to you as a training manager?

: :	: 7 :	: :	: 2 :	: :
NO USE AT ALL	OF VERY LITTLE USE	MODERATELY USEFUL	VERY USEFUL	EXTREMELY USEFUL

Comments?
MODIA did not tell us anything we didn't already know.

16. How useful was the cost information on the alternative course designs to you as a manager?

: 2 :	: 5 :	: 1 :	: 1 :	: :
NO USE AT ALL	OF VERY LITTLE USE	MODERATELY USEFUL	VERY USEFUL	EXTREMELY USEFUL

Comments?

17. **Do you feel the simulation of alternative designs could be of value to you as a course manager?**

5 **Yes**   4 No

18. **In what ways would you use information provided by the cost model?**

19. **What was** the cost relationship of the baseline course to the **alternatives** you asked the ISD team members to plan on MODIA?

Alternative 1:
1 a. Much more expensive
1 b. More expensive
4 c. About the same
3 = No response
___ d. Less expensive
___ e. much less expensive

Alternative 2:
___ a. Much more expensive
1 b. More expensive
3 c. About the same
___ d. Less expensive
3 = No response
___ e. Much less expensive

Comments or explanations?

20. Were the alternative course designs workable -- that is, did they conform to Air Training Command and Technical School policy?

1 a. Completely workable

4 b. Workable with minor changes

2 c. Somewhat unworkable - major changes required

2 = No response ____ d. Totally unworkable

Comments?

21. In your opninion, how should MODIA be used (a short sentence or two)?

22. In your opinion, who should use MODIA (Specify "who" at each organizational level, i.e., training evaluation, plans, operations etc. -- You can specify NONE or MORE THAN ONE)?

Technical School Personnel (Center Level): 4

Technical Training Group personnel: 3

Branch Personnel: 3 course planners/curricula

Hq ATC personnel: 4

2 - No one should use it.

23. For each organizational level you checked, please, in a sentence or two, explain why they would use MODIA?

Technical School:

Technical Training Group:

Branch:

Hq ATC:

24. What do you think would be the role of those organizations you indicated in using MODIA?

25. Please list any additional comments you care to make about your experience with MODIA, its usefulness, or any suggestions you may have for improving the system.

- Increase program limits on teaching formats and agents.
- Increase number of learning events
- Improve cost model to permit insertion of other course cost. One weak area encountered is in expendable supplies. Our training courses use materials that are costly in supporting performance training in the laboratories.
- The system must be expanded to be worthwhile.
- The programs need to be expanded.

DISASSOCIATED UTILITY OF MORIBUND BRAINS

by

CDR C. F. Meredith, USCG

Thank you, Captain FERGUSON, for the opportunity to present my paper this evening. After the paper I presented last year in San Antonio, I was classified as a standard deviant. Subsequently, I was advised to get closer to NORM; the only NORM's I know are non-standard deviants.

The major thrust of my research has been in the area of disassociated utility of moribund brains, acronymically,

DUMB

In reality the full title of my paper is Disassociated Utility of Moribund Brains in Stratified Higher Intellectual Technology.

Unfortunately, I was unable to locate an appropriate acronym in the U. S. Government Catalog of Standard Acronyms on which to base my paper and subsequent research, therefore, I have reverted to the symplistic form DUMB.

Disassociated Utility of Moribund Brains in Stratified Higher Intellectual Technology

This study degenerated from a self-conceptualized realization that the parathetical basis for psychomotorial and congenital evaluative processes, derived from replications of the cause-defect continuum in U. S. military training is, in itself, a process of debilitating obfuscatory criterion-referenced retrograde directed systemization which has as its propitiary conclusion a higher order of lesser inactivity in the non-results-oriented result of out-processing of human resources, or, if you will, why so many military trainees are revolting.

To encapsulate, in the initiatory process of learner-referenced behavior modification, symbolism is employed in varying degrees in representative relationships. For example, observe this series of symbols:

I I I I I I

Each of these inter-related digitally displayed symbols have a cross-related definition, if you will, an object, an entity, a being unified essence of quantifiable quality. In laymen's terminology, apples and oranges. Through an interactive process involving psycho-motor applications, these symbols can be interposed and juxtaposed in a variety of arrays to produce a specific differential resultant, terminally speaking.

I shall now depict in graphic form through an interaction of cylinder-form calcium-based substance and a vertically-oriented green-hued non-organic slate object, mis-termed a blackboard, how these symbols are most commonly presented to the learning inputee: 1 + 1 =

The substantive nature of these symbols has been non-imperically transformed. Yet, and herein lies the crux of my considerations, the arrangement of these data has not led to a predicted conclusion and if we co-locate and additional non-relative symbol

$$1 + 1 = ?$$

our perception also communicates a significant discertitude.

My research to date has led only to a preliminary conclusion. By a random selection of one symbol from the population of similar data and applying the aforementioned methodology, I have found that the digital array can approach content validity.

$$1 + 1 = 2$$

Traditionalists in our field have supported my findings (OG 4200 B.C., Einstein 1909 A.D.). On the other hand, those who have subscribed to the precepts of stratified higher intellectual technology have aritculated interrogatism. I would be remiss in this paper if I failed to replicate the differentiations. But, before I graphically display the argument against my approach, I shall reiterate synthetically my self-propogated fear that if the research of the stratified higher intellectual technologists reaches an unnatural conclusion which is the usual result, disassociated utility of moribund brains (or dumbness) will be the terminal orientation.

Their non-articulated objection in sum ostensibly stems from a perceived non-utilization of inherantly dichotomous symbolism leading to and causatory of the disassociated properties of my partially stratified bias-oriented selection of the digital data. Their methodology suggests the elimination of the chance-level symbol

$$?$$

whereby one is restricted to the imposition of only one additive similar symbol, and further suggests the selection of three similar symbols thereby resulting in this analogous if illogical formulation:

$$1 + 1 = N$$

In conclusion, I am gratified to state that my research reached termination in the pre-data gathering stage and fortunately will not be published. I will be happy to question any of your answers after the conclusion of this evening's program.

REPORT OF STEERING COMMITTEE

and

GENERAL BUSINESS MEETING (1978)

1. The Steering Committee recommended and the membership approved changes to the by-laws which redefined a quorum of the Steering Committee and instructed the Secretary to solicit nominations for the Harry H. Greer Award.

2. A description of the Harry H. Greer Award and its recipients will be appended to the by-laws.

3. The German Armed Forces Association and the German Armed Forces Psychological Service Research Institute were accepted as primary members of the Steering Committee.

4. A list of the primary membership of the Steering Committee will be appended to the by-laws.

5. The coordinating agencies of the next four annual conferences will be:

1979	Naval Personnel Research and Development Center (San Diego)
1980	Canadian Forces Personnel Applied Research Unit (Toronto)
1981	Army Individual Training Evaluation Directorate (Ft. Eustis)
1982	Naval Education and Training Program Development Center (Pensacola)

BY-LAWS OF THE MILITARY TESTING ASSOCIATION*

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

---

*As approved at the 1978 General Meeting of the Association 2 Nov 78, Oklahoma City, Oklahoma

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

STEERING COMMITTEE MEMBERS
# of the
MILITARY TESTING ASSOCIATION

1. Naval Personnel Research and Development Center
2. Naval Education and Training Program Development Center
3. Army Research Institute
4. Air Force Human Resources Laboratory
5. Air Force Occupational Measurement Center
6. Army Individual Training Evaluation Directorate
7. U. S. Coast Guard Institute
8. Canadian Forces Personnel Applied Research Unit
9. Canadian Forces Directorate for Manpower Occupational Structures
10. Royal Australian Air Force Evaluation Division
11. German Armed Forces Association
12. German Armed Forces Psychological Services Research Institute

HARRY H. GREER AWARD

The Military Testing Association is an outgrowth of an informal meeting of representatives of the various armed forces testing agencies in 1958. The meeting was held at the suggestion (and through the personal coordination) of CAPT Harry H. GREER, USN, Commanding Officer of the Naval Examining Center. Thus, CAPT GREER was the "founder" of the Military Testing Association. In 1962, an award in his name was created to recognize significant lasting contributions to the Association while exemplifying the ideals of the Association and its founder.

The five recipients of the award since 1962 are:

1962	CAPT Harry H. GREER, USN
1970	COL J. M. McLANATHAN, USAF
1974	MR. C. J. MacALUSO, Naval Examining Center
1977	DR. W. J. MOONAN, Naval Personnel Research and Development Center
1977	MR. J. A. BURT, U. S. Coast Guard Institute

INDEX OF AUTHORS

AND

LIST OF CONFEREES

PAGE

ADAMICK, Daniel R.
    USAOCCS, Aberdeen Proving Ground, Attn: ATSL-TD-TD, Maryland 21005
ADAMS, MAJ Jerome Ph.D.
    3098-B Stony Lonesome, West Point, NY 10996
  Paper presented: "Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Decription of Leader Behavior. 434
ADAMS, William
    Chief of Naval Education & Training, Pensacola, FL 32508
ADKINS, Homer
    Chief of Naval Education & Training, NAS Pensacola, Florida 32508
ALLMAN, CAPT Thomas S.
    USAF Squadron Officers School/Chief, Standardization Division, SOS/EDVS, Maxwell AFB, AL 36112
ANDERSON, Kermit B.
    1408 Spruce, Norman, OK 73069
ANSBRO, Thomas M.
    CDG CNET N-5 Bldg 679, NAVAIRSTA, Pensacola, FL 32508
  Paper presented: "Using the Computer to Build the Task Inventory 263
ANZELMO, CAPT Ralph H. (USMC)
    HQMC Office of Manpower Utilization (MPU), Quantico, Virginia 22134
ASA-DORIAN, Paul V.
    Fleet Anti-Submarine Warfare Training Center, San Diego California 92147
AUMENT, John (USAF)
    443D TCHTS/QUV, Altus AFB, OK 73521
AVERSANO, Dr. Francis M.
    ATTSC-IT-TD, US Army Training Support Center, Fort Eustis, VA 23604
  Paper presented: "Task Analysis: Destination or Journey" . . 199
BABIN, Ms. Nehama
    Army Research Institute, 5001 Eisenhower Ave., Alexandria, Virginia 22333
  Paper presented: "Differential Field Assignment Patterns for Male and Female Soldiers" 396

PAGE

BARAN, Harry A.
34 Old Yellowsprings Rd., Fairborn, OH 45324
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602

BARBER, Herbert F.
US Army Research Institute Field Unit, P.O. Box 3122, Ft. Leavenworth, KS 66027
Paper presented: "Critical Performances of Battalion Command Groups" 1264

BARRON, Clovis J.
USN Naval Education & Training Program, Development Center - Code PD10, Pensacola, FL 32509

BEEL, C. D.
Naval Manpower Utilization Unit, HMS Vernon, Portsmouth PO1 3ER, Hampshire, England
Paper presented: "Execution of Large Occupational Analysis of the Royal Navy's Operations Branch". 112

BEGLAND, CAPT Robert R.
Training Development Institute, USA TRADOC, 123 Tabb Lane, Tabb, VA 23602
Paper presented: "How Do You Buy 'Good Design": An Examination of the Army's TEC Program". 1098

BELL, 1LT Steven J., MSC
Training Evaluation Division, DTDE, AHS, Superintendent, Academy of Health Sciences, USA, ATTN: HSA-TEC, Fort Sam Houston, TX 78234

BENNETT, CPT Oscar D.
Academy of Health Sciences, ATTN: HSA-TIP, Fort Sam Houston, TX 78234

BERGMANN, Joseph A.
AFHRL/ORA, Brooks AFB, TX 78235
Paper presented: "Female Utilization in Non-Traditional Areas". 444

BERNSTEIN, LCDR David M.
HQ, USCG Reserve Training Division, 400 7th Street, SW, Washington, DC 20590

BILLS, CAPT Conrad G.
USAF Occupational Measurement Center, USAFOMC/OMDC, Lackland AFB, TX 78236
Paper presented: "Evaluation of Computer-Derived Test Outlines Using Conventional Test Outlines as a Criterion Reference During Test Development Projects" 976

BIRDSALL, Walter W.
Naval Education & Training Program Development Center, Ellyson, Pensacola, FL 32509

PAGE

BLANKENSHIP, Constance
Navy Personnel Research and Development Center, San Diego, CA 92152
Paper presented: "The Premature Attrition of Navy Female Enlistees". 420

BODRON, LCDR Donald E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

BOLDT, R. F.
Educational Testing Service, Rosedale Rd., Princeton, New Jersey 08541
Paper presented: "Some Implications of Commercial Test Normings for Mobilization Surveys". 633

BONETTE, Cedella J.
USA Military Personnel Center, DAPC-MSP-D, 200 Stovall St., Alexandria, VA 22332
Paper presented: "General Overview and Initial Findings of the Project on Job Satisfaction and Retention of US Army Enlisted Personnel" 75

BOONE, Dr. James O.
FAA Civil Aeromedical Institute, Mike Monroney Aeronautical Center, P.O. Box 25082, Oklahoma City, Oklahoma 73125
Paper presented: "A New Procedure to Make Maximum Use of Available Information When Correcting Correlations for Restriction in Range Due to Selection". 906

BOSSHARDT, Michael J.
Personnel Decisions Research Institute, 2415 Foshay Tower, Minneapolis, MN 55402
Paper presented: "Content Validation of Class A School Curricula in the Coast Guard" 1107

BOTHWELL, Cheryl
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

BOWER, CAPT Frederick B. Jr.
USAF Occupational Measurement Center, Lackland AFB Texas 78236
Paper presented: "The Stability Over Time of Air Force Enlisted Career Ladders as Observed in Occupational Survey Reports" 228

BOWNAS, David A.
Personnel Decisions Research Institute, 2415 Foshay Tower, Minneapolis, MN 55402
Paper presented: "Content Validation of Class A School Curricula in the Coast Guard" 1107

PAGE

BOWSER, Samuel E.
5900 Lake Murray Boulevard, LaMesa, CA 92041

BRADNER, Dr. Cleveland Jr.
PD5, Naval Education & Training Program Development
Center, Ellyson, Pensacola, FL 32509

BREWER, ENS David B.
Reserve Training Division, USCG Headquarters,
Washington, DC 20590

BUCK, DR. C. W.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

BURNS, Darla J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

BURNS, Dr. Eugene M.
USA Military Personnel Center, ATTN: DAPC-MSP-SM,
200 Stovall St., Alexandria, VA 22332
Paper presented: "Evaluating the Army Occupational Survey Program Methodology: Answer Booklets, Questionnaire Length, and Population Coverage". 51

BURT, John A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

BURTCH, Lloyd D.
Air Force Human Resources Laboratory, Brooks AFB,
Texas 78235
Paper presented: "A Methodology to Evaluate the Aptitude Requirements of Air Force Jobs" 1012

BURTON, LTJG Richard T.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

BYHAM, W.
Development Dimensions, Inc., Pittsburg, PA
Paper presented: "Development of the Army ROTC Management Simulation Program and Instructors' Orientation Program". 1091

CARGILL, Bonnie K.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

CARPENTER, Dr. James B.
KENTRON International, Inc., 14023 Rocky Pine Woods,
San Antonio, TX 78249

CARRAWAY, Jay
Naval Education & Training Program Development Center,
Box 212A Rt. 4, Pensacola, FL 32504

CARTER, LCDR Clinton, W.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

PAGE

CASTELNOVO, Anthony E.
US Army Research Institute, P.O. Box 3066, Ft. Still, Oklahoma 73503
Paper presented: "Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" . 1091
Paper presented: "Prediction of Field Artillery Officer Performance". 839
CHAGALIS, CPT George P.
Academy Health Sciences, Room 247, Ft. Sam Houston, Texas 78233
CHASE, LT Philip K.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
CHRISTAL, Raymond E.
AFHRL/ORA, Brooks AFB, TX 78235
Paper presented: "Female Utilization in Non-Traditional Areas 444
CONN, Barbara A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
COOK, ENS Deborah J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
CORY, Charles H.
Navy Personnel Research & Development Center, San Diego, California 92152
Paper presented: "Assessment Center Variables as Predictors of On-Job Performance Characteristics". 761
COWAN, Douglas K.
AFHRL, Brooks AFB, TX 78235
Paper presented: "Civilian Ground Safety Officer Job and Training Requirements Survey" 16
CRIMMINS, CWO2 James H.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
CRONIN, ENS Michael J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
CUMMINGS, CAPT William H.
ATC Technology Applications Center, Lackland AFB, TX 78236
Paper presented: "Job Performance of USAF Bypassed Specialists". 724

PAGE

CUNNINGHAM, J. W.
North Carolina State University, 2205 Hillsborough, Raleigh, NC 27607
Paper presented: "Determining the Training Requirements of United States Coast Guard Warrant and Commissioned Officer Billets". 28

CZUCHRY, Andrew J.
Dynamics Research Corporation, Wilmington, MA
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602

DAPRA, R. A.
Development Dimensions, Inc., Pittsburgh, PA
Paper presented: "Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" . 1091

DAVILA, LTJG Robert E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

DAVIS, D. Douglass
Chief of Naval Education & Training (CNET), Naval Air Station, Pensacola, FL 32508
Paper presented: "Data Base To Determination of Training Content: A Manageable Solution". 258

DELONEY, Rebecca
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

DeVRIES, Philip B.
McDonnell Douglas Astronautics Co., P.O. Box 516, St. Louis, MO 63166
Paper presented: "Methods for Collecting and Analyzing Task Analysis Data" 314

DeVRIES, LCDR Richard L.
RESGRU SW Hbr 01-88804, Box 147, RFD 1, Rockland, ME 04841

DICKINSON, Richard W.
Computer Programming & Statistical Analysis, Occupational Research Program - Industrial Engineering, Texas A&M University, College Station, TX 77843

DIETERLY, Duncan L.
USAF Human Resources Laboratory, Wright-Patterson AFB, Ohio 45433
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602

DITULLIA, Paul
USAF Occupational Measurement Center, Lackland AFB, TX 78236

DOORLEY, Richard D.
USA Military Personnel Center, 200 Stovall St., Rm 1S23, Alexandria, VA 22332

PAGE

DOW, Dr. Andrew N.
USNETPDC - Ellyson, Pensacola, FL 32509
Paper presented: "Objective Evaluation of Correspondence Course Items" 1027

DREW, LT Richard
Officer in Charge, Central Test Site for PTEP, NAVGMS
Virginia Beach, VA 23461

DREWES, D. W.
North Caroline State University, 2205 Hillsborough,
Raleigh, NC 27607
Paper presented: "Determining the Training Requirements of United States Coast Guard Warrant and Commissioned Officer Billets". 28

DRISKILL, Dr. Walter E.
USAF Occupational Measurement Center/OMYO, Lackland AFB,
Texas 78236
Paper presented: "Four Fundamental Criteria for Describing the Tasks of an Occupational Specialty" 204
Paper presented: "The Stability Over Time of Air Force Enlisted Career Ladders as Observed in Occupational Survey Reports" 228

DUFFY, Paul C.
Marine Corps Institute, Marine Barracks 8th & I Sts.,
P.O. Box 1775, Washington, DC 20013

DURHAM, MAJ Charles V.
Evaluation Branch, Academic Instructor School, USAF,
AIRFOS, EDV, Maxwell AFB, AL 36112

DYER, Dr. Frederick N.
Army Research Institute Field Unit, P.O. Box 2086,
Ft. Benning, GA 31905
Paper presented: "Using an Assessment Center to Predict Leadership Course Performance of Army Officers and NCOs . 779

EARLES, James A.
AFHRL/PES, Brooks AFB, TX 78235
Paper presented: "The Content Issue in Performance Appraisal Ratings". 508

EASTMAN, Robert F.
US Army Research Institute Field Unit, P.O. Box 476,
Ft. Rucker, AL 36362
Paper presented: "Validity of Associate Ratings of Performance Potential by Army Aviators" 823

ELLIS, Dr. John A.
Navy Personnel Research & Development Center, Code 304,
San Diego, CA 92152
Paper presented: "The Instructional Quality Inventory: Introduction and Overview". 1138

PAGE

ELLIS, CAPT R. T.
Canadian Forces Personnel Applied Research Institute,
4900 Yonge St., Willowdale, Ontario, Canada

ESCHENBRENNER, Dr. A. John
McDonnell Douglass Astronautics Co., P.O. Box 516,
St. Louis, MO 63166
Paper presented: "Methods for Collecting and Analyzing
Task Analysis Data" 314

ESLICK, CWO4 David W.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

EVANS, Ermon M.
Chief of Naval Technical Training, (CNTECHTRA),
704 W. Sherrod, Covington, TN 38019

FARNSWORTH, LT Barry A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

FARRIS, John C.
Data-Design Laboratories, L5 Koger, P.O. Box 12773,
Norfolk, VA 23502

FERGUSON, CAPT. J. E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

FINE, Dr. Sidney A.
Advanced Research Resources Org., 4330 East West Highway,
Bethesda, MD 20014
Paper presented: "Analysis of Heavy Equipment Operator
Jobs" 734

FISCHL, Dr. M. A.
317 Rexburg Ave., Fort Washington, MD 20022
Paper presented: "Measuring the Military Base Population
of the 1980's". 640

FOGLE, Charles C.
FAA Airman Examinations AFS-593, AERO Center, Will
Rogers Field, OK

FOLEY, Paul P.
3965 Aqua Dulce Blvd., Spring Valley, VA 92077

FORMAN, Ima R.
USCG Institute, P.O. Substation 18, Oklahoma City, Ok
73169

FREY, Dr. Robert L. Jr.
USCG (G-P-1/2/62), Washington, DC 20590

GAFNEY, LT Edward J. III
Strategis Systems Project Office, Crystal City Mall 3,
Washington, DC 20376

PAGE

GENTNER, CAPT Frank C.
USAF Occupational Measurement Center/OMYO, Lackland AFB, Texas 78236
Paper presented: "Four Fundamental Criteria for Describing the Tasks of an Occupational Specialty" 204
GEORGE, Dr. Clay E.
Dept. of Psychology, Texas Tech University, Lubbock, TX 79409
Paper presented: "An Analysis of the OE Concept and Suggested Improvements" 942
GERBER, Dr. J. E. Jr.
HQ US Army Forces Command, KTTN AFPR-PSE, Ft. McPherson, Georgia 30330
GILBERT, Dr. Arthur C. F.
US Army Research Institute, 5001 Eisenhower Ave., Alexandria, VA 22333
Paper presented: "Prediction of Field Artillery Officer Performance". 839
Paper presented: "Predictive Utility of the Officer Evaluation Battery (OEB)" 753
Paper presented: "Quality of ROTC Accessions to the Army Officer Corps". 488
GIORGIA, M. Joyce
Air Force Human Resources Lab, Brooks AFB, TX 78235
GOCLOWSKI, John C.
Dynamics Research Corporation, Wilmington, MA
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602
GOLDMAN, Dr. Lawrence A.
USA Military Personnel Center, DAPC-MSP-D, 200 Stovall St., Alexandria, VA 22332
Paper presented: "General Overview and Initial Findings of the Project on Job Satisfaction and Retention of U.S. Army Enlisted Personnel" 75
GOODGAME, Doug
Occupational Research Program, Industrial Engineer Dept., Texas A&M University, College Station, TX 77801
Paper presented: "Scheduling Formal School Training to Maximize Cost Effectiveness". 286
GOODY, Kenneth
AFHRL, Brooks AFB, TX 78235
Paper presented: "Benchmark Scales for Collecting Task Training Factor Data" 556
GORDON, Mr. M. Meriwether
AF ROTC, AFROTC/ACME, Maxwell AFB, AL 36112
Paper presented: "Weighted Selection System for AFROTC Applicants--Perspective After Second Year of Use" 566

PAGE

GOULS, Dr. R. Bruce
AFHRL/PES, Brooks AFB, TX 78235
GRAHAM, Dr. William W. Jr.
MEPCON, MEPCT-P, Bldg. 83, Ft. Sheridan IL 60037
Paper presented: "Development of a Mobilization Population Inventory Using Existing ASVAB Data Banks". 645
GRIMM, Richard
PD-10, NETPDC Ellyson, Bldg. 922, Pensacola, FL 32509
GROETKEN, LTC David L.
Chief, Analysis Div. Directorate of EVAC, USA Field Artillery School, Ft. Sill, OK 73503
GROVER, Martha S.
Defense Intelligence School, Washington, DC 20374
GUERREIN, Joseph H.
USA Infantry School, SFTD, Directorate of Training, Ft. Benning, GA 31905
HALADYNA, Tom
Oregon College of Education, Monmouth, OR 97361
Paper presented: "The Emergence of an Item-Writing Technology" 1035
HALTRECHT, Dr. Ed
Personnel Research, Ontario Hydro (H2-D17), 700 University Ave., Toronto, Ontario M5G 1X6
HANLON, John P.
Ft. Devens, MA 01433
HASSALL, LCDR James L.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
HASSEN, John E.
Code N5B2, Chief of Naval Education & Training Support, Bldg. 997, Ellyson Field, Pensacola, FL 32509
HAWRYSH, CDR Fred J.
Directorate of Military Occupational Structures, Canadian Forces, National Defense Headquarters, Ottawa, Ontario, Canada K1A 0K2
HEJL, CWO4 L. E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
HENDERSON, Robert G.
Defense Language Institute Foreign Language Center, ATTN: ATLF-TD-JS, Presidio of Monterey, CA 93940
Paper presented: "The Defense Language Aptitude Battery (DLAB)" 574
HENN, LT COL Manfred
MOD Germany, Ministry of Defense - Armed Forces Staff I3, Postfach 1328, 5300 Bonn 1, W. Germany

PAGE

HICKS, Dr. Jack M.
6827 Old Chesterbrook Rd., McLean, VA 22101
Paper presented: "Leader Sex, Leader Descriptions of Own Behavior, and Subordinates Description of Leader Behavior" 434
Chairman, Symposium: "Methodology for Mobilization Population Inventory" 632

HILLIGOSS, Richard E.
Army Research Institute Field Unit, P.O. Box 2086,
Ft. Benning, GA 31905
Paper presented: "Using an Assessment Center to Predict Leadership Course Performance of Army Officers and NCOs". 779

HOUTZ, John C.
USA Recruiting (USARCASP-E), Ft. Sheridan, IL 60037

HOWARD, Dr. Charles W.
805 Cortijo, El Paso, TX 79912
Paper presented: "Methodology for Evaluating Operator Performance on Tactical Operational Simulator/Trainers" . 1255

HUNTER, John E.
Michigan State University, East Lansing, MI 48823
Paper presented: "The Impact of Valid Selection Procedures on Workforce Productivity". 677
Paper presented: "Test of a New Model of Validity Generalization: Results for Tests Used in Clerical Selection" . 856

JACKSON, Alvaline B.
3060D Mower Court, Ft. Mead, MD 20755
Paper presented: "Evaluation of Intelligence Producing Capability of Selected Combat Arms Units" 1205

JACKSON, LT COL David K.
AFROTC/ACME, Maxwell AFB, AL 36112
Paper presented: "Weighted Selection System for AFROTC Applicants--Perspective After Second Year of Use" 566

JACKSON, William L.
Directorate of Training Developments, Training Analysis
& Design Division, Ft. Rucker, AL 36362

JENKINS, William J.
US Army, Redstone Arsenal, AL 35481

JENNINGS, Alan E.
FAA CAMI, AAC118, P.O. Box 25082, Oklahoma City, OK
73125
Paper presented: "A Method to Evaluate Performance Reliability of Individual Subjects". 933

JENNINGS, Margarette C.
Advanced Research Resources Organization, 4330 East West
Highway, Bethesda, MD 20014
Paper presented: "Analysis of Heavy Equipment Operator Jobs" 734

PAGE

JOHNSON, LT David G.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
JOHNSON, Dorothy
Corry Station, Cryptologic Dept., Pensacola, FL 32503
JOHNSON, Kirk A.
Naval Personnel Research & Development Center, 6392 Amberly St., San Diego, CA 92120
JOHNSON, Robert N.
USA Administration Center (DTD), Ft. Benjamin Harrison, Indiana 46216
Paper presented: "Design of Machine Scorable 'Hands On' Performance Tests in a Paper and Pencil Mode" 1161
JONES, Dr. Jean
Army Research Institute for the Behavioral & Social Sciences, HQ TCATA (PERI-OH), Ft. Hood, TX 76544
JONES, Karen N.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
JONES, Dr. Todd
R&D US Coast Guard (G-DSA-1/TP44), Washington DC 20590
KAHN, Dr. Arthur
Westinghouse Defense & Electronic Systems Center, P.O. Box 746 (M.S. 440), Baltimore, MD 21203
Paper presented: "Experimental Evaluation of a High Technology Training Program". 1116
KAPLAN, Dr. Ira T.
Training Development, US Army Research Institute Field Unit, P.O. Box 3122, Ft. Leavenworth, KS 66027
Paper presented: "Critical Performances of Battalion Command Groups" 1264
KEATES, CAPT W. E.
Staff Officer Analysis, Air Command Headquarters, Westwin, Manitoba, Canada R2R OTO
Paper presented: Aircrew Training Research - Project ACTIVE" 1068
KEETH, James B.
USAF Occupation Measurement Center, Lackland AFB, TX 78236
KINNISON, Henry L.
Dept. of Psychology, Texas Tech University, Lubbock, Texas 79409
Paper presented: "An Analysis of the OE Concept and Suggested Improvements" 942
KINTOP, Constance
MGR Personnel Services, Minneapolis Personnel Dept., 312-3RD Ave. South, Minneapolis, MN 55415

PAGE

KNAUP, Peggy A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

KNERR, Dr. C. Mazie
US Army Research Institute, 5001 Eisenhower Ave., ATTN: PERI-OU, Alexandria, VA 22333
Paper presented: "An Application of Tactical Engagement Simulation for Unit Proficiency Measurement". 1316

KNIGHT, Patricia
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

KOHL, Delbert E.
Marine Corps Institute, Marine Barracks, Box 1775, Washington, DC 20013

KOSKI, LT John D.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

KRAIN, Dr. Burton F.
US Civil Service Commission, Intergovernmental Personnel Programs Division, 230 S. Dearborn, Chicago, IL 60604

KRIETEMEYER, CDR George E.
USCG Aviation Technical Training Center, Elizabeth City, North Carolina 27909

KUENZ Dr. Marjorie A.
Naval Health Sciences, Education & Training Command, National Naval Med. Center, Bethesda, MD 20014
Paper presented: "Systematic Instructional Validation Through Testing". 275

KUHNLE, CDR Robert L.
Leadership Program Staff, USCG Reserve Training Center, Yorktown, VA 23690

LAABS, G. J.
Navy Personnel Research & Development Center, San Diego, California 92152
Paper presented: "Performance Test Objectivity: Comparison of Interrater Reliabilities of Three Observation Formats". 831

LAMBRECHT, Marvin W.
1515 S. Jefferson Davis Hwy, Arlington, VA 22202

LANTERMAN, Richard S.
US Coast Guard, 400 7th St., SW, Washington, DC 20590
Paper presented: "Content Validation of Class A School Curricula in the Coast Guard" 1107

LEECH, LT COL Carl A.
Canadian Forces Directorate of Military Occupational Structures, National Defense Headquarters, 101 Colonel By Drive, Ottawa, Ontario Canada K1A 0K2

PAGE

LEFROY, MAJ Dal
CF PARU, Suite 600, North York Govt. of Canada Bldg.,
Toronto, Ontario, Canada M2N 6B7
LEGER, Marie
US Army Research Institute, 5001 Eisenhower Ave.,
Alexandria, VA 22333
Paper presented: "Validity of Associate Ratings of
Performance Potential by Army Aviators" 823
LEHMAN, LT Stanley E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
LEWIS, Dr. John R.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
LEWIS, Dr. Mary A.
AAC-118 FAA/CAMI, P.O. Box 25082, Oklahoma City, OK 73125
Paper presented: "A Comparison of Three Models for Determining Test Fairness" 919
LINCOLN, John O.
Defense Language Institute, English Language Center,
Lackland AFB, TX 78236
LINDSEY, Shellie
6403 E. 16th, Anchorage, AK 99504
LINNON, CDR J. L.
USCG Training Center, Governors Island, New York, NY 10004
LIU, Georgina
Army Education Center, Ft. Ord, CA 93941
LOFASO, Anthony J.
Dynamics Research Corp., Wilmington, MA
Paper presented: "PAM: A Methodology for Predicting
Air Force Personnel Availability" 602
LONG, James L.
CNET (Code N-531), NAS Pensacola, FL 32508
LOTZ, George Jr.
4713 NW 59th Terrace, Oklahoma City, OK 73122
LOWE, Muriel
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
MARCO, Ruth Ann
McDonnell Douglas Astronautics Co., P.O. Box 516,
St. Louis, MO 63166
Paper presented: "Methodology for Selection and Training
of Artillery Forward Observers Job Analysis" 324
MARTIN, J. Thomas Jr.
Data Design Laboratories, 15 Koger Executive Center,
Suite 140, Norfolk, VA 23502
Paper presented: "A Comparison of Two Criterion-Referenced
Scoring Procedures for an Answer-Until-Correct, Multiple-
Choice Performance Test". 938

PAGE

MARTIN, LTJG Thomas J.
USCG Headquarters (G-PMR-5), 400 7th St., SW,
Washington, DC 20590
MASSEY, CAPT Randy H.
AFHRL/PES, Brooks AFB, TX 78235
Paper presented: "The Content Issue in Performance
Appraisal Ratings". 508
MATHEWS, John J.
AFHRL, Brooks AFB, TX 78235
Paper presented: "Prediction of Reading Grade Levels of
Service Applicants from Armed Services Vocational
Aptitude Battery (ASVAB)" 494
MEEK, Patricia A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
MEREDITH, CDR Carlton F.
USCG AVTRACEN, Mobile, AL 36608 1373
MEREDITH, Dr. John B. Jr.
Data Design Laboratories, P.O. Box 12773, 15 Koger,
Norfolk, VA 23502
Paper presented: "A Comparison of Two Criterion-Referenced
Scoring Procedures for an Answer-Until-Correct, Multiple-
Choice Performance Test". 938
MERRILL, M. David
Courseware, Inc. San Diego, CA
Paper presented: "The Instructional Quality Inventory:
Introduction and Overview". 1138
MESSICK, Vernon D.
NETPDC, Ellyson, Pensacola, FL 32509
MESSURA, CWO3 Ronald A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
METTY, CWO4 Cleo F.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
MILLIGAN, Dr. John R.
US Army Research Institute, P.O. Box 3066, Ft. Sill,
Oklahoma 73503
Paper presented: "A Learning-Receptive State as Induced by
an Auditory Signal or Frequency Pulse". 1181
Paper presented: Observer Self-Location Ability and its
Relationship to Cognitive Orientation Skills" 333
MINTER, CDR Richard W.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
MITCHELL, LT COL Jimmy L.
USAFOMC/ONY, Stop 100, Lackland AFB, TX 78236
Paper presented: Differential Responses on Alternately
Anchored Job Rating Scales" 525

PAGE

MOBLEY, Amelia E.
USCG 400 7th St., SW, Washington, DC 20590
MOCHARNUK, Dr. John B.
McDonnell Douglas Astronautics Co. P.O. Box 516,
St. Louis, MO 63166
Paper presented: "Methodology for Selection and Training of
Artillery Forward Observers Job Analysis" 324
MONROE, CWO3 Larry N.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
MONTELMERLO, Dr. Melvin D.
ATTSC-IT-TD, US Army Training Support Center,
Ft. Eustis, VA 23604
Paper presented: "Task Analysis: Destination or Journey" 199
MULDROW, Tressie W.
Resound Research & Development Center, US Civil Service
Commission, 1900 E St., NW, Washington, DC 20415
Paper presented: "The Impact of Valid Selection Procedures
on Workforce Productivity:. 677
MULLANE, CAPT Thomas F.
Service School Command, Naval Training Center, Orlando,
Florida 32813
MULLINS, C. J.
AFHRL/PES, Brooks AFB, TX 78235
Paper presented: "The Content Issue in Performance
Appraisal Ratings". 508
MURPHY, John W.
USAIA Test Design Coordinator, Ft. Benjamin Harrison,
Indiana 46218
MUSSIA, Stephen J.
Manager, Personnel Research & Eval., Minneapolis Personnel
Dept., 312 3rd Ave. South, Minneapolis, MN 55415
MYERS, David C.
Advanced Research Resources Organization, 4330 East West
Hwy., Bethesda, MD 20014
Paper presented: "Analysis of Heavy Equipment Operator
Jobs" 734
McCLINTOCK, Dr. William R.
PD-10 Navy Education & Training Program Development
Center, Bldg. 922, NETPCD, Ellyson, Pensacola, Fl 32509
McCOY, Linda A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
McDANIELS, CWO2 Donald M.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169
McINTOSH, Virgil M.
AF Extension Course Institute (EDV), Gunter AFS, AL 36118

PAGE

McIVER, LT COL Werner W.
Headquarters, USMC, Office of Manpower Utilization,
Quantico, VA 22134

McKENZIE, Robert C.
Resound Research & Development Center, US Civil Service
Commission, 1900 E St., NW, Washington, DC 20415
Paper presented: "The Impact of Valid Selection Procedures
on Workforce Productivity". 677

McLANATHAN, COL Frank L.
St. Mary's University, San Antonio, TX 78284

McVAY, Kenneth W.
USA Missile & Munitions Center & School, ATTN: ATSK-TD-PM,
Redstone Arsenal, AL 35809

NEFF, Edwin F.
Assistant Chief Training & Education, US Coast Guard,
400 7th St., SW, Washington, DC 20590

NELSON, Oliver
USAF/ATC Randolph AFB, TX 78148

NODDIN, Ernest M.
Submarine Medical Research Laboratory, Box 900, Submarine
Base, Groton, CT 06340

NOVAK, Frank J.
Naval Education & Training Program Development Center,
Classified Instructional Material Division, (PD-9)
Bldg 942, Ellyson, Pensacola, FL 32509

NOWLIN, Debra L.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

NUGENT, William A.
Navy Personnel Research & Development Center, ATTN: Code
9309, San Diego, CA 92152
Paper presented: "Performance Test Objectivity: Comparison
of Interrater Reliabilities of Three Observation Formats" 831

O'CONNELL, Joseph
Police Training, M.L.E.O.T.C., 7426 North Canal Rd.,
Lansing, MI 48913

O'LEARY, Dr. Brian S.
U.S. Civil Service Commission, 1900 E Street, NW,
Washington, DC 20415
Paper presented: "Construct Validity". 849

OLIVER, Dr. L. W.
Army Research Institute, 5001 Eisenhower Ave.,
Alexandria, VA 22333
Paper presented: "Differential Field Assignment Patterns
for Male and Female Soldiers" 396

OLIVO, CAPT John
USAF Occupation Measurement Center, Lackland AFB, TX
78236
Paper presented: "The Use of Job Satisfaction Data in the
Occupational Survey Program". 65

PAGE

OLSON, Howard C.
Advanced Research Resources Org., 4330 East West Hwy., Bethesda, MD 20014
Paper presented: "Analysis of Heavy Equipment Operator Jobs" 734

ORRISON, CPT Stephen L.
Director of Training Development/USAIS, ATTN: ATSH-I-U-TDD, Ft. Benning, GA 31905

OSBORNE, James E.
NAVEDTRA PRODEVCEN (PD-3), Ellyson AFB, Pensacola, FL 32509

OVERTON, Deborah J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

PACKER, CWO4 Harold R.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

PARKS, LT Alton J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

PARSONS, Tom
KENTRON International, Inc., P.O. Box 35417, Brooks AFB, San Antonio, TX 78235

PASS, Dr. John J.
Navy Personnel Research & Development Center, San Diego, California 92152
Paper presented: "Sample Size and Stability of Task Analysis Inventory Response Scales" 537

PASTENE, ENS Charles R.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

PATTERSON, CAPT Gary K.
USAF Sheppard AFB, Wichita Falls, TX 76308

PEARLMEN, Kenneth
US Civil Service Commission, 1900 E St., NW, Washington, DC 20415
Paper presented: "Test of a New Model of Validity Generalization: Results for Tests Used in Clerical Selection" . 856

PESKOE, Stuart E.
Dynamics Research Corporation, Wilminton, MA
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602

PETERSON, CWO3 Phillip M.
USCG, 10 Philanne Dr., Norwich, CT 06360

PHALEN, William J.
AFHRL/ORA, Brooks AFB, TX 78235
Paper presented: "The Development of a Technique for Using Occupational Survey Data to Construct and Weight Computer-Derived Test Outlines for Air Force Specialty Knowledge Tests (SKTs)" 949

PAGE

PHILLIPS, Fredric F.
Dynamics Research Corporation, Wilmington, MA
Paper presented: "PAM: A Methodology for Predicting Air Force Personnel Availability" 602

PORTER, George V. Jr.
Director, Cadet Exams & Records, USAF Academy, CO 80840

POTTER, LT Earl H. III
US Coast Guard Academy, New London, CT 06320

PROVENMIRE, H. K.
USCG Aviation Training Center, Bates Field, Mobile, Alabama 36608

POWELL, Ladonna A.
USCG Inistitute, P.O. Substation 18, Oklahoma City, OK 73169

PUZICHA, Dr. Klaus
Regierunsdirektor Bei, Dezernat Wehrpsychologie IM, Streitkrafteamt, ABT. I

QUICK, Bob J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

RAMPTON, LCOL Glenn M.
Canadian Forces Personnel Applied Research Unit, 4900 Yonge St., Willowdale, Ontario
Paper presented: "A Strategy for Task Analysis and Criterion Definition Based on Multidimensional Scaling" . 132

RAY, MAJ W. D.
Directorate of Evaluation, USAMPS/TS, Ft. McClellan, Alabama 36205

RECKASE, Dr. Mark D.
University of Missouri, 4 Hill Hall, Columbia, MO 65211
Paper presented: "A Generalization of Sequential Analysis to Decision Making with Tailored Testing" 994

REINHARDT, LCDR William H.
US Navy Occupational Development & Analysis Center (NODAC), Bldg 150, Washington Navy Yard (ANACOSTIA), Washington, DC 20374

RENEAU, ENS Lee
USCG Training Center, Cape May, NJ 08204

RICHARDS, Robert E.
The Pennsylvania State University, State College, PA
Paper presented: "The Instructional Quality Inventory: Introduction and Overview". 1138

ROBERTS, Fred C.
Naval Health Sciences, Education & Training Command, National Naval Med. Center, Bethesda, MD 20014
Paper presented: "Systematic Instructional Validation Through Testing". 275

PAGE

ROBERTSON, D. W.
Navy Personnel Research & Development Center, San Diego, CA 92152
Paper presented: "Sample Size and Stability of Task Analysis Inventory Response Scales" 537

ROID, Gale
Teaching Research Division, Oregon State System of Higher Education, Monmouth, OR 97361
Paper presented: "The Emergence of an Item-Writing Technology" 1035

ROOT, Robert T.
US Army Research Institute, 5001 Eisenhower Ave., Alexandria, VA 22333
Paper presented: An application of Tactical Engagement Simulation for Unit Proficiency Measurement". 1316

RUBRIGHT, Earl
80th MTC, 556 Valleywood Dr., Millers Ville, MD
Paper presented: "Evaluation of Intelligence Producing Capability of Selected Combat Arms Units" 1205

RUCK, Hendrick W.
AFHRL/OR, Brooks AFB, TX 78235
Paper presented: "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development". 242
Paper presented: "Methods for Collecting and Analyzing Task Analysis Data" 314
Paper presented: "Methods for Determining Safety Training Priorities for Job Tasks" 296
Paper presented: "Obstacles to and Incentives for Standardization of Task Analysis Procedures" 188
Paper presented: "A Technique for Selecting Electronic Specialties for Consolidation". 385

RUMSEY, M. G.
US Army Research Institute, 5001 Eisenhower Dr., Alexandria, Virginia 22333
Paper presented: "Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" . 1091

RUX, George V.
MEPCON, MEPCT-P, Bldg. 83, Ft. Sheridan IL 60037
Paper presented: "Development of a Mobilization Population Inventory Using Existing ASVAB Data Banks". 645

SANDS, William A.
Navy Personnel Research & Development Center (Code 310), San Diego, CA 92152
Paper presented: "Computer Assisted Reference Locator (CARL) System: An Overview". 470

PAGE

SARGENT, Mildred L.
Naval Education & Training Program Development Center, Ellyson Field, Pensacola, FL 32509

SCANLAND, Dr. Dorothy
US Naval Education & Training Command (Code N-5), Pensacola, FL 32508

SCANLAND, Dr. Worth
US Naval Education & Training Command (Code N-5), Pensacola, FL 32508

SCHIEMANN, William A.
Project Manager, AT&T, Rm. 6126F2, 295 N. Maple Ave., Basking Ridge, NJ 07920

SCHMIDT, Frank L.
Resound Research & Development Center, US Civil Service Commission, 1900 E St., NW, Washington, DC 20415
Paper presented: "The Impact of Valid Selection Procedures on Workforce Productivity". 677
Paper presented: "Test of a New Model of Validity Generalization: Results for Tests Used in Clerical Selection" . 856

SCHWARTZ, CWO4 John E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

SCOTT, LT Lynn M.
AFHRL/ORA, Brooks AFB, TX 78235

SEIBEL, David
D. E. Siebel & Assoc. LTD (Canadian Forces), #1609 1275 Richmond Rd., Ottawa, Ontario K2B 8E3

SELLMAN, MAJ Wayne S.
Air Force Manpower and Personnel Center, Randolph AFB, Texas 78418
Paper presented: "Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB)" 494

SEUBERLICH, COL Hans-Erich
Chairman Army Section DBwV (Federal Armed Forces Association), Sudstrabe 123, 5300 Bonn 2, W. Germany
Paper presented: "Strain by Prolonged Duty Hours and Problems as to Mobility of Soldiers - As Seen by Federal Armed Forces Association" 463

SHIPLEY, Brian D. Jr.
US Army Research Institute Field Unit, P.O. Box 476, PERI-OA, Ft. Rucker, AL 36362
Paper presented: "Complexity of Flight Path Data as an Index of Skill in Piloting Performances from a Flight Simulator Based Job-Sample Test". 1193
Paper presented: "Learning Aptitude, Error Tolerance, and Achievement Level as Factors of Performance in a Visual-Tracking Task". 1220

PAGE

SHIVELY, Albert E.
Naval Education & Training Program Development Center, Ellyson, Pensacola, FL 32509
SHOEN, William R.
Service School Command, Naval Training Center, Orlando, FL 32813
SILVERSTEIN, Jerome H.
Defense Language Institute, English Language Center, Lackland AFB, TX 78236
SIMS, Dr. Bill
Center for Naval Analyses, 1401 Wilson Blvd. Arlington, Virginia 22209
SKOFSTAD, Dennis
Aviation Technical Training Center, Elizabeth City, North Carolina 27909
SMITH, Dr. Bea H.
Naval Amphibious School, Coronado, San Diego, CA 92155
SMITH, H. Wayne
Dept. of Psychology, Texas Tech University, Lubbock, TX 79409
Paper presented: "An Analysis of the OE Concept and Suggested Improvements" 942
SOLOMON, Elberta
PD5, Naval Education & Training Program Development Center, Ellyson, Pensacola, FL 32509
SPRAGUE, LT Chester M.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
STAMM, LTJG James A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
STEFFEN Dale A.
Electronics Division, Denver Research Institute, University of Denver, P.O. Box 10127, Denver, CO 80210
Paper presented: "Evaluation of Troubleshooting Simulator" 1249
STEPHENSON, Donald P.
Staff & Faculty Division, Office of DAC for EEL Tech, USAARMS, Ft. Knox, KY
STEPHENSON, Dr. Robert W.
AF Human Resources Laboratory, Brooks AFB, TX 78235
Paper presented: "Obstacles to and Incentives for Standardization of Task Analysis Procedures" 188
STERLING, Martha E.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
STEWART, RADM W. H.
Chief, Office of Personnel, USCG, 400 7th St., SW, Washington, DC 20590
Paper presented (Keynote Address): "Quality of Life" . . . xi

PAGE

STIMATZ, LT J. Anthony
Dept. of Mathematics, USCG Academy, New London, CT 06320
SVEJKOVSKY, Mary L.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
TAKAHASHI, Terry
Defense Intelligence School, Washington, DC 20374
TALLEY, John W.
SWT Division, ATSK-TD-AD-A, Bldg 3342, USAMMCS, Redstone Arsenal, AL 35809
TARTELL, J. S.
Academy of Health Sciences, Ft. Sam Houston, TX 78234
Paper presented: "Job Analysis in the US Army Medical Training Environment" 354
TAYLOR, Donald F.
CG Research & Development, 2366 Antiqua Ct. Reston, VA 22091
TAYLOR, CAPT Ronald L.
Extension School, Education Center, Marine Corps Development & Education Command, Quantico, VA 22134
TEMPLEMAN, Max
Chief, Education Branch, US Army Support Command, DPCA, USASCH, Ft. Shafter, HI 96858
THAIN, John W.
Defense Language Institute, Presidio of Monterey, Monterey, CA 93940
Paper presented: "Monte Carlo Computer Programs for Simulating Selection Decisions from Personnel Tests". . . 586
THEW, Michael C.
AFHRL/SMAZ, Brooks AFB, TX 78235
Paper presented: "CODAP: A New Modular Approach to Occupational Analysis". 362
THOMAS, Patricia J.
Navy Personnel Research and Development Center, San Diego, California 92152
Paper presented: "The Premature Attrition Rate of Navy Female Enlistees" 420
THOMPSON, CDR George J.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
THOMPSON, Nancy
AFHRL/OR, Brooks AFB, San Antonio, TX 78229
Paper presented: "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development". 242
Paper presented: "Methods for Determining Safety Training Priorities for Job Tasks" 296

PAGE

THOMSON, David C.
609 Sunhaven Dr., San Antonio, TX 78239
Paper presented: "Benchmark Scales for Collecting Task Training Factor Data" 556
Paper presented: "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development". 242
THURING, LT Allen R.
USCG Reserve, 943 N. Liberty St., Arlington, VA 22205
TRATTNER, Marvin H.
US Civil Service Commission, 1900 E Street, NW, Washington, DC 20415
Chairman, Symposium: "Innovative Test Validation Strategies" 848
Paper presented: "Synthetic Validity". 879
VALENTINE, Dr. Lonnie D. Jr.
6205 Rue Francois, San Antonio, TX 78238
Paper presented: "Air Force Experience with PROJECT TALENT" 671
Paper presented: "Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB)". 494
VAN NOSTRAND, SALLY J.
US Army Research Institute, 5001 Eisenhower Ave., Alexandria, VA 22333
Paper presented: "Occupational Analysis for Field Grade Army Officers". 373
VAUGHAN, CAPT David S.
ATC Technology Applications Center, Lackland AFB, TX 78236
Paper presented: "Job Performance of USAF Bypassed Specialists". 724
Paper presented: "Two Applications of Occupational Survey Data in Making Training Decisions". 213
VOORHEES, Phyllis L.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 731269
WALDKOETTER, Dr. Raymond O.
US Army Research Institute, P.O. Box 3066, Ft. Sill, Oklahoma 73503
Paper presented: "A Learning-Receptive State as Induced by an Auditory Signal or Frequency Pulse". 1181
Paper presented: "Observer Self-Location Ability and Its Relationship to Cognitive Orientation Skills" 333
Paper presented: "Prediction of Field Artillery Officer Performance". 839

PAGE

WALLIS, M. Reid
Richard A. Gibboney Associates, Kensington, MD
Paper presented: "Occupation Analysis for Field Grade Army Officers" 373
WARM, Thomas A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169
Paper presented: "A Primer of Item Response Theory". . . . 884
WEBER, CAPT Elena J.
USAF Occupational Measurement Center, Lackland AFB, TX
Paper presented: "The Use of Job Satisfaction Data in the Occupational Survey Program". 65
WEHR, CDR Robert H.
USCG Aviation Training Center, Bates Field, Mobile, AL 36608
WEHRENBERG, STC Stephen B.
USCG Reserve Training Center (OGLAMS), Yorktown, VA 23690
WEISSMULLER, Johnny J.
AFHRL/SMAZ, Brooks AFB, TX 78235
Paper presented: "CODAP: A New Modular Approach to Occupational Analysis". 362
WELDON, Dr. John I.
US Army Training and Doctrine Command, Alexandria, VA
Paper presented: "Quality of ROTC Accessions to the Army Officer Corps". 488
WELDON, Roland L.
Course Development Division, US Army Aviation Center, Ft. Rucker, AL 36362
WELLINS, Dr. Richard S.
US Army Research Institute, 5001 Eisenhower Ave., Alexandria, VA 22333
Paper presented: "Development of the Army ROTC Management Simulation Program and Instructors' Orientation Course" . 1091
Paper presented: "Quality of ROTC Accessions to the Army Officer Corps". 488
WELSH, CAPT John R. Jr.
3307 School SQ., Lackland AFB, TX 78236
Paper presented: "Evaluation of the MODIA Planning System" 1335
WERNER, MAJ Gerald C.
Dept. of Army Individual Training, HQ DA, ATTN: DAMO-TRI, The Pentagon, Washington, DC 20310
WEST, Anita S.
Denver Research Institute, University of Denver, P.O. Box 10127, Denver, CO 80210
Paper presented: "Evaluation of Troubleshooting Simulator" 1249

PAGE

WHITE, Jonathan
Police Training, M.L.E.O.T.C., 7426 N. Canal Rd., Lansing, MI 48913

WILCOVE, Gerry L.
Navy Personnel Research & Development Center, San Diego, California 92152
Paper presented: "The Premature Attrition of Navy Female Enlistees". 420

WILLHOITE, CAPT Robert R.
The National Bank of Commerce, Altus, OK 73521

WILLIAMS, Rayburn A.
Chief of Naval Education & Training N-53, Naval Air Station, Pensacola, FL 32506

WILLIAMSON, Sharon A.
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

WILLING, Richard
USCG Institute, P.O. Substation 18, Oklahoma City, OK 73169

WINKLER, Edward B.
Human Resources Management School, Naval Air Station, Millington, TN

WINN, Francis J. Jr.
Medical Unit, USCG Support Center, Governors Island, New York, NY 10004

WISKOFF, Dr. Martin
Navy Personnel Research & Development Center, 5151 Dixel Dr., San Diego, CA 92115

WITTMAN, LTC Clarence E.
US Army Directorate of Evaluation, US Army Field Artillery School, Ft. Sill, OK 73503

WOOD, Norman D.
The Pennsylvania State University, State College, PA
Paper presented: "The Instructional Quality Inventory: Introduction and Overview". 1138

WORD, LTC Larry E.
US Army Training Support Center, Ft. Eustis, VA
Paper presented: "An Application of Tactical Engagement Simulation for Unit Proficiency Measurement". 1316

WORSTINE, Darrell A.
USA Military Personnel Center, DAPC-MSP-D, 200 Stovall St., Alexandria, VA 22332
Paper presented: "General Overview and Initial Findings of the Project on Job Satisfaction and Retention of U.S. Army Enlisted Personnel". 75

PAGE

WULFECK, Wallace H. II
Navy Personnel Research & Development Center, Code 304,
San Diego, CA 92152
Paper presented: "The Instructional Quality Inventory:
Introduction and Overview". 1138

YOUNG, LT Larry C.
USCG Institute, P.O. Substation 18, Oklahoma City, OK
73169

ADDITIONAL CONFEREES

BLAND, CDR R. D.
USCG Training Center, Cape May, N.J. 08204
CARLSON, Robert R.
Commandant (G-PTE), U.S. Coast Guard, Washington, D.C. 20590
CHIPPENDALE, Joan
TRACEN Governors Island, NY 10004
CRUICKSHANK, James G.
Commandant (G-PTE), U.S. Coast Guard, Washington, D.C. 20590
DONOHOE, CAPT L.V.
USCG Training Center, Cape May, N.J. 08204
GARCIA, LT Rebecca M.
Commander (r), 11th Coast Guard District, 400 Oceangate,
Long Beach, CA 90822
GREENFIELD, LCDR J. T.
Commander (r), 11th Coast Guard District, 400 Oceangate,
Long Beach, CA 90822
JOYCE, LCDR E. P.
U.S. Coast Guard Academy, New London, CT 06320
PALESE, Robert
TRACEN Governors Island, NY 10004
SANOK, CDR Gregory J.
USCG TRACEN, Gov't Island, Alameda, CA 94501
THRALL, CAPT F. E.
TRACEN Governor's Island, NY 10004